REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
Nov 95

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Chemometric Analysis Of Multimode Fluorescence Data obtained With a Pulsed Tunable Laser

5. FUNDING NUMBERS

6. AUTHOR(S)
Mark Hilary Van Benthem

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFIT Student Attending:
North Dakota State University

8. PERFORMING ORGANIZATION REPORT NUMBER
96-018D

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
DEPARTMENT OF THE AIR FORCE
AFIT/CI
2950 P STEET, BLDG 125
WRIGHT-PATTERSON AFB OH 45433-7765

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release IAW 190-1
Distribution Unlimited
BRIAN D. GAUTHIER, MSgt, USAF
Chief Administration

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS

15. NUMBER OF PAGES
160

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT

18. SECURITY CLASSIFICATION OF THIS PAGE

19. SECURITY CLASSIFICATION OF ABSTRACT

20. LIMITATION OF ABSTRACT

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C -	Contract	**PR** -	Project
G -	Grant	**TA** -	Task
PE -	Program Element	**WU** -	Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

★U.S.GPO:1990-0-273-271

CHEMOMETRIC ANALYSIS OF MULTIMODE FLUORESCENCE DATA OBTAINED WITH A PULSED TUNABLE LASER

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Mark Hilary Van Benthem

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Chemistry

November 1995

Fargo, North Dakota

ABSTRACT

Van Benthem, Mark Hilary, Ph.D., Department of Chemistry, College of Science and Mathematics, North Dakota State University, November 1995. Chemometric Analysis of Multimode Fluorescence Data Obtained With a Pulsed Tunable Laser. Adviser: Dr. Gregory D. Gillispie.

This research evaluated the capabilities of various chemometric methods for analysis of three-mode fluorescence data. Data were collected using pulsed Nd:YAG laser and pulsed Nd:YAG laser-pumped dye laser excitation and monochromator-PMT-digital oscilloscope detection. This apparatus produced data in the form of nanosecond scale time decay profiles at numerous emission wavelengths generating a wavelength-time matrix (WTM). Third-order data were produced by varying analyte concentrations or changing excitation wavelength to produce a time-resolved excitation-emission matrix (TREEM). Trilinear decomposition (TLD) and global analysis methods were applied to a WTM-concentration 3-array and a TREEM. TLD methods and linear discriminant analysis and classification were performed on highly complicated data in the form of WTMs of various fuels.

The three-mode data were decomposed with an eigenanalysis-based procedure (EBP); three-mode alternating least squares (3M-ALS), also known as PARAFAC; and three-mode nonnegative alternating least squares (3M-NNALS).

Various rank estimation procedures were evaluated for the two-mode and three-mode data. The efficacy of classification and clustering algorithms applied to reduced forms of three-mode fuel fluorescence data was also demonstrated.

Analyses of the WTM-concentration 3-array illustrate the capabilities of various rank estimation and profile extraction techniques with low-rank data. This four-component

WTM-concentration 3-array was easily decomposed with TLD methods. 3M-ALS and 3M-NNALS offered a slight improvement to the EBP result. 3M-NNALS offered a minor computation time improvement over 3M-ALS.

A TREEM of a four-component solution of fluorene, naphthalene, carbazole, and phenanthrene in water was measured. TLD was difficult because of extensive spectral and time-mode overlap. The EBP and 3M-ALS were unable to provide realistic factor profiles. 3M-NNALS produced very good results, generating recognizable factors in one-tenth of the time required by 3M-ALS.

WTMs of fuels on soil matrices were analyzed. TLD using an EBP was not successful, yielding complex eigenvectors or chemically meaningless factors. 3M-ALS performed better than the EBP, but it produced multiple factor degeneracies and took a great deal of computation time. 3M-NNALS performed the best of the three TLD procedures. Computation time was at least two orders of magnitude faster than 3M-ALS.

ACKNOWLEDGMENTS

I dedicate this work to my family: my darling wife, Sue; my beloved son, Max; and my precious daughter, Heidi. Sue and Max understood, inspired, and loved me throughout this challenging time; and Heidi brought joy into all of our lives when she took her first breath in January of this year.

I extend my sincerest thanks to my good friend and adviser, Professor Gregory D. Gillispie, for all he has done for me through the years. I hope his new and future enterprises bring him success and happiness.

For all of their help and understanding, I thank the members of Professor Gillispie's research group: my good friend, Dr. Orven Swenson, who was always willing to proofread and tutor; Randy St. Germain, who collected so much data; Daniel Halkin, who taught me how to use a computer; Dr. Jim Burt; Dr. William Cooper; Dr. Michael Dvorak; Dr. Ben Mitchell; Lance Delabarre; Joe Klingfus; Roxane Meidinger; Gregory Neubauer; Greg Oswald; and Mr. Harlan Isensee.

I thank the members of my graduate committee for all of their help and understanding; the faculty, staff, and graduate students in the Department of Chemistry; and the staffs of the NDSU Libraries and Information Technology Services.

Finally, I acknowledge and thank the United States Air Force for this opportunity, the people of the Air Force Institute of Technology for their cooperation, and Brigadier General Howard J. "Foot" Ingersoll for making it happen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

Abbreviation/Acronym	Meaning
µm	micrometer (10^{-6} meters)
3M-ALS	three-mode alternating least squares
3M-NNALS	three-mode nonnegative alternating least squares
ADC	analog-to-digital converter
AFA	abstract factor analysis
ALS	alternating least squares
CCD	charge coupled device
CFD	constant fraction discriminator
DF	diesel fuel
DFM	diesel fuel marine
DoD	United States Department of Defense
DSO	digital storage oscilloscope
EBP	eigenanalysis based procedure
EEFA	excitation-emission frequency array
EEM	excitation-emission matrix
EFA	evolutionary factor analysis
EPA	Environmental Protection Agency
fs	femtosecond (10^{-15} seconds)
GS/s	gigasamples per second
HCA	hierarchical cluster analysis
ILS	inverse least squares
KHz	kilohertz
m	meter
MCA	multichannel analyzer
MHz	megahertz
MLO	multiple local optima
MLS	multivariate least squares
NBRL	non-bilinear rank annihilation
Nd:YAG	neodymium-doped yttrium aluminum garnet
NIR	near infrared
nm	nanometers (10^{-9} meters)
NNALS	nonnegative alternating least squares
NNLS	nonnegative least squares
ns	nanosecond (10^{-9} seconds)
PAH	polycyclic aromatic hydrocarbon
PARAFAC	parallel factor analysis
PCR	principal component regression
PF	principal factor
PFA	principal factor analysis

PLS	partial least squares
PMT	photomultiplier tube
ppb	parts per billion by weight (μg/kg)
ppm	parts per million by weight (mg/kg)
ps	picosecond (10^{-12} seconds)
RAFA	rank annihilation factor analysis
RBL	residual bilinearization
ROST	Rapid Optical Screening Tool
S/N	signal-to-noise ratio
SCAPS	Site Characterization and Analysis Penetrometer System
SMCR	self-modeling curve resolution
SVD	singular value decomposition
TAC	time-to-amplitude converter
TCSPC	time-correlated single-photon counting
TFA	target factor analysis
TLD	trilinear decomposition
TREEM	time-resolved excitation-emission matrix
UG	unleaded gasoline
WTM	wavelength-time matrix

1. INTRODUCTION

Chemical analysis provides information about the composition of a material (which might be groundwater in a well, soil under a filling station, an ingot of a metal alloy, etc.) to address a concern someone has about the material.[1] For example, is water in the well safe to drink? Is there underground petroleum contamination in the property for sale? Can this alloy be used in a turbine? Answering the fundamental question posed to the analyst requires determination of one or more chemical parameters. The desired parameters are sometimes qualitative in nature, i.e., identities of chemical components (what heavy metals are in the water, petroleum products in the soil, strategic metals in the alloy), but more often quantitative information, i.e., the amounts of chemical components (how much lead is in the groundwater, gasoline in the soil, or chromium in the alloy, etc.) is desired. The job of the analytical chemist is to obtain the best information in the necessary form to answer these questions. It is in the interest of both the analytical chemist and the client to obtain accurate data in a fast, inexpensive, and minimally invasive fashion.

Traditional analysis methods require collection of representative samples from the bulk material for subsequent laboratory analysis. Laboratory analysis methods provide great specificity at the expense of slow speed and the requirement of extensive sample preparation and tedious data analysis. In addition, the precision and the accuracy of the data are often limited by the quality of the sampling,[2] which may not have been performed by the chemist;[3] for environmental projects, the chemist, in fact, rarely has any connection with the sampling. Analytical techniques that allow measurement of bulk materials in real time and in situ would be a windfall to analytical chemistry. Fiber-optic spectroscopy provides such utility in a number of situations.

Fiber-optic spectroscopy can be extremely valuable in situations where sampling is very difficult, dangerous, or time critical. It may also be indicated when analysis involves obtaining information about a remote or otherwise inaccessible location, for example, spectroscopic monitoring of jet engine exhaust in flight. One may also use it to monitor multiple stages of an industrial process simultaneously. Safety while performing analysis is another issue that may dictate the use of fiber optics. Spectroscopic analysis of explosives or radioactive materials are examples of analyses that may be more safely performed at a distance using fiber optics.

Near-infrared (NIR) absorbance spectroscopy performed over fiber optics has been practiced for many years[4] and has reached a relatively advanced stage of commercial maturity. NIR instruments are used for industrial process monitoring and medical diagnostics and in environmental chemistry. In applications for which quantitative information is required for species at high concentrations, it can be a superior analysis method, saving time and money. Because samples are not taken, however, special consideration must be given to interferences, i.e., absorbance by species other than the target analytes.

Traditional analytical methods provide opportunities to eliminate sample interferences before the actual data collection. Great effort is expended to eliminate interferences since they can be a dominant source of error. If the analyte response cannot be separated from that of the interference, one must model the influence of the interferences on the analyte response. Fiber-optic spectroscopy and other in situ techniques, which generally do not allow sample manipulation, are especially susceptible to interferences and matrix effects. The burden of detecting and treating the interferences falls on the quality of the spectroscopic data and, ultimately, on the data analysis.

To correctly predict the concentration of a spectroscopically active species, all spectroscopically active species present in the unknown must be included in the calibration model.[5] If there are interfering species whose contributions were not accounted for in the calibration procedure, the concentration determination will be biased.[6] For chemically complex systems, as are often encountered in environmental analyses, interferences can be a difficult problem to overcome.

Multivariate analysis of NIR spectroscopic data represents a highly refined approach to modeling absorbance data and is the source of the technique's popularity and utility. However, because NIR (and other absorbance techniques) are limited to a single data *mode*,[*] viz., wavelength, their ability to predict unknown parameters in the presence of interferences not included in the calibration model is limited. *Multimode* data, of *order*[†] greater than that available in absorbance, can allow calibration and analysis even if there are interferences not modeled in the calibration.[5,7]

Fluorescence spectroscopy is readily adapted to multimode techniques.[8] Moreover, it is several orders of magnitude more sensitive than absorbance[9] and, unlike absorbance

[*] In many works, the term *dimension* is used in place of the term *mode*, which will be used here. The term mode indicates the functionality employed to acquire data. For example, wavelength is the mode utilized in absorbance spectroscopy, while charge to mass ratio is the mode of mass spectral data, time is the mode of gas chromatographic data, etc. Mode is explained in greater detail in Appendix B.

[†] The meaning of the term order is discussed in greater detail in Chapter 3; however, consider it to represent the number of *modes* used to collect data.

techniques, is easily applied to highly scattering matrices like soils. By measuring the fluorescence emission spectrum at different excitation wavelengths (or equivalently, measuring the excitation spectrum at different emission wavelengths), one can generate a matrix of data commonly referred to as an excitation-emission matrix (EEM). The two *modes* of an EEM are the excitation wavelength and the emission wavelength. Similarly, one can measure the fluorescence emission at various emission or excitation wavelengths as a function of time following pulsed excitation to generate a wavelength-time matrix (WTM). A series of excitation-emission matrices obtained at various time delays following pulsed excitation represents three data modes. Given the higher order techniques potentially available and the overall versatility of the technique, fluorescence spectroscopy promises to be a superior spectroscopic analysis method in many circumstances.

Unfortunately, fluorescence spectroscopy has an inherently lower signal-to-noise ratio (S/N) than absorbance spectroscopy.[9] The ramifications here are lower precision measurements, hence, lower precision parameter determination (e.g., concentration). Of course, in situations that would otherwise lead to large sampling uncertainty or great risk associated with sampling, the precision trade-offs of a fiber-optic fluorescence measurement may be acceptable. Again, one must consider the type of information desired of the measurement (the data quality objectives).

If the analysis goal is quantitative information, the analysis requires a calibration step followed by an unknown estimation step. The presumption in this case is that one knows the identity of the species present and desires the concentrations. However, when the goal of the analysis is qualitative information, one is faced with a more thorny problem;

especially if the material has many unknown components with similar spectra and there are multiple matrix effects, as in environmental analysis. The parameters of interest now become the relative proportions of the spectral intensities as a function of wavelength, rather than their absolute magnitudes, i.e., the spectra of the contaminants. The data must be decomposed to yield the individual spectra of the components.

The presence of noise in the spectral data can make decomposition difficult, leading to ambiguous results.[10] Here, too, the multimode data provided by fluorescence compensate greatly in spite of the lower S/N compared to absorbance spectroscopy. Once again, the analyst is faced with trade-offs between S/N and sensitivity and higher data order. In fact, three-mode (and higher mode) data with linear independence between the modes (so-called trilinear data) produce a unique decomposition if some other mild assumptions are met.[11,12] Simulations of decomposition of trilinear data with various levels of noise demonstrate that analysis of fluorescence spectra is very encouraging.[10]

Given the many potential advantages of fluorescence spectroscopy, its increasing use in the chemical analysis of complex systems is not surprising. Environmental analysis has been impacted immensely by the marriage of fiber optics and fluorescence.[13] The EPA's list of the 20 most significant hazardous substances includes several fluorescent species.[14] Most common fuels contain significant levels of fluorescent compounds, specifically benzene and its single aromatic ring derivatives, naphthalene and substituted naphthalenes, and other polycyclic aromatic hydrocarbons (PAHs).[15-17] Fuels also contain high levels of aliphatic hydrocarbons, which are not fluorescent (but could be determined by NIR). However, in terms of toxicity and carcinogenicity, the aliphatic compounds do not pose nearly the threat that the aromatic species do.[18] It is fortunate, then, that fluorescence selectively detects the species of greatest concern.

The collection of truly representative environmental samples from a suspected contaminated site can be *extremely challenging* and *expensive*,[19] particularly in the case of subsurface samples. In the standard approach, the areas of interest must be drilled or core sampled. The installation cost of a single monitoring well can exceed $10,000, and this money is wasted if the well is not placed in an appropriate location.

Once a well has been installed, the samples must be collected, packaged, shipped, stored, and analyzed judiciously so they are not corrupted. Loss of sample integrity could occur at any or several of these steps. Fuel-contaminated soil and water samples that contain easily lost volatile and semi-volatile components must be handled with great care. The elapsed time between sampling and availability of the chemical analysis report may be several weeks. To make matters worse, after the drilling or core sampling is finished, a substantial conduit remains for transport of the contaminants to other soil strata. Finally, analysis methods proscribed by EPA to extract contaminants from soil[20] involve the use of solvents which themselves are hazardous materials. Methods that would allow results in real time with lower costs are certainly desirable. A promising method to measure fuel contaminants in soil combines fiber-optic fluorescence with cone penetrometry.

Cone penetrometry is a method of determining geophysical parameters of subsurface soil. It was developed primarily to characterize geology of areas during oil, gas, and mineral exploration. It employs a hydraulic press mounted in a very heavy (≈20 tons) truck to drive sections of stainless steel pipe (≈5 cm diameter) into the earth. The pipe contains various sensors to measure geotechnical parameters such as sleeve friction and tip resistance; these parameters form the basis of a soil-type characterization scheme. A sapphire window in the side of the pipe close to the tip of the probe allows optical access to the soil. Fiber optics leading to and from the window allow direct fluorescence examination of the soil. A new device has been developed specifically for this type of spectroscopy. The Rapid Optical Screening Tool (ROST™, Dakota Technologies, Inc., Fargo, ND) allows collection of a fluorescence time-resolved excitation-emission matrix (TREEM) of petroleum-contaminated soil using fiber-optic transmission of excitation and emission.

The ROST™ utilizes a frequency doubled, Nd:YAG laser pumped-dye laser to generate ultraviolet laser excitation which can be passed into an optical fiber. The detector incorporates a scanning monochromator, photomultiplier tube (PMT), and digital storage oscilloscope (DSO). The output of the DSO is channeled into a personal computer (PC) and stored for processing. The coupling of ROST™ with the cone penetrometer makes a powerful tool that can quickly produce multimode spectra of contaminated soil to rapidly characterize a contaminated site. Compared to traditional methods of chemical analysis, cone penetrometry with fiber-optic detection is more cost effective and may be more accurate.

The rationale for pursuing cone penetrometry-based fluorescence analysis is compelling. Its strength, in addition to the factors mentioned above, lies in the wealth of information obtained during an analysis. In addition, after performing cone penetrometry, the hole produced can be grouted as the probe is removed. This minimizes the spread of any hazardous material, unlike monitoring wells.

Rarely do sites contain a single compound (e.g., benzene). The fuels and residue from other energy-producing activities (e.g., manufactured gas plants) represent very complex mixtures of fluorescing species. Consequently, the spectra of the contaminated soils are usually quite complex. The higher order inherent in fluorescence data may allow more reliable and complete chemical analysis than do single-mode techniques.

Data obtained from cone penetrometry-based fluorescence is extensive. Consider measuring fluorescence intensity as a function of excitation wavelength, emission wavelength, fluorescence decay time, and probe depth at various locations on a site. It is easy to see how large the data sets can be. It is not practical to perform analysis, other than gross generalizations, on such enormous data sets without the use of a computer. Fortunately, there is a relatively young branch of chemistry designed to deal with data analysis of large and complicated data sets: chemometrics.

Chemometrics has been defined as using chemical principles and mathematical and statistical methods to interpret and predict chemical data.[21] Much of chemometrics is an offshoot of developments in applied statistics used by psychometricians and econometricians to analyze data related to human behavior and the performance of markets, respectively.[22] The general idea is to take a set of data pertaining to a group of subjects and draw inferences about the nature of the group. For example, a psychologist might want to analyze infant behavior in relation to different stimuli to learn about language development. It is also valuable to process the data to find a predictive method for subjects within or outside of the group. Economists, and almost everyone else, would like to be able to accurately predict when a stock price will go up and when it will go down. It may be apparent to the reader that some humans have exceptional methods of "data analysis" without the use of computers and complex algorithms. Given the proper training and/or experience, people can be very effective at detecting trends, recognizing patterns, estimating proportions, and predicting events. Certainly, there are very astute market speculators who make a very good living simply using their "feel" about the market, just as a skilled spectroscopist can identify a compound in a mixture and accurately estimate its concentration by briefly examining a spectrum.

Analytical chemistry, though, demands much more of analysis than a rough guess, no matter how accurate. Chemometrics provides the analytical chemist with numerical tools capable of dealing with large and complicated sets of data and reducing them to a useful and desired form.

Data reduction is the hallmark of a good chemometric analysis routine. From large sets of fluorescence data, one seeks the nature and quantity of the species giving rise to the fluorescence. These may be from contaminants in soil and groundwater or some other source. The objective here, and the subject of this thesis, was to examine chemometric methods that could assist an analyst in determining the identities of luminescent species in solutions and mixtures or the identity of a composite mixture (e.g., a fuel) using fluorescence data obtained from the ROST™ or a similar instrument.

Chapter 2 of this work reviews various properties of fluorescence with particular attention paid to those aspects of fluorescence exploited by the ROST™. Chapter 3 addresses the chemometric methods used to decompose ROST™ acquired data. The principles of factor analysis are discussed to identify or quantify unknowns. Methods of separating collections of data into groups using pattern recognition and classification schemes are the subject of Chapter 4. Results of the research conducted and complementary discussion are presented in Chapter 5. Chapter 6 incorporates conclusions about this research and recommendations for future work. For those not familiar with multivariate analysis and matrices, Appendix A compiles a selection of definitions and explanations of concepts covered throughout the thesis. Appendixes B and C also contain background material relevant to the subject of higher order factor analysis.

2. FLUORESCENCE REVIEW

2.1 Fluorescence Principles

Fluorescence in organic molecules is often related to the presence of delocalized π electrons,[23] most notably in aromatic hydrocarbons and their derivatives. In the discussion of fluorescence principles, we shall use the language and nomenclature conventions relevant to the fluorescence of aromatic molecules.

Absorption of a photon by a molecule in the ground electronic state, S_0, promotes the molecule into one of the states, S_n, of the excited electronic state manifold. The integrated strength of the electronic transition is governed by an integral involving the electronic wavefunctions, while the shape of the excitation spectrum is determined by the Franck-Condon factors. In general, electronically excited molecules in the condensed phase rapidly relax ($\sim 10^{-12}$ s) to a Boltzmann distribution of vibrational levels in the first excited singlet state, S_1 via the processes of internal conversion ($S_2 \rightarrow S_1$) and vibrational relaxation ($v' >> 0 \rightarrow v' \sim 0$). The absence of appreciable fluorescence from states higher in energy than S_1 is known as Kasha's rule.[24]

Deactivation from the S_1 state occurs through a variety of mechanisms. The radiative S_1 to S_0 deactivation process is referred to as fluorescence. Others processes include internal conversion ($S_1 \rightarrow S_0$), external conversion (energy transfer and quenching), and intersystem crossing (singlet to triplet conversion). The role of the Franck-Condon factors in determining the shape of the emission spectrum is similar to the way that they affect the excitation spectrum. After returning to S_0, the molecule vibrationally relaxes to a

Boltzmann distribution of vibrational levels. These collective phenomena have two important consequences: (1) for a given molecule, the shape of the fluorescence emission spectrum (i.e., the relative intensity distribution as a function of wavelength) is independent of the excitation wavelength; and (2) the shape of the fluorescence excitation spectrum is independent of the emission wavelength monitored during its acquisition.

The depopulation rate of the excited state also has important consequences. The differential equation for the disappearance of S_1 after photoexcitation is

$$-\frac{\mathrm{d}[\mathrm{S}_1]}{\mathrm{dt}} = k_r[\mathrm{S}_1] + k_{nr}[\mathrm{S}_1] + k_p[\mathrm{S}_1] + k_q[\mathrm{S}_1][\mathrm{Q}], \tag{2.1}$$

where $[\mathrm{S}_1]$ is the concentration of molecules in the first excited state, k_r is the first-order rate constant for radiative transition, k_{nr} is the first-order rate constant for non-radiative transitions combining internal conversion and intersystem crossing, k_p is the first-order rate constant for unimolecular photochemistry, k_q is the second-order rate constant for quenching, and $[\mathrm{Q}]$ is the concentration of a quenching species. The rate constants in (2.1) are usually combined into a pseudo-first-order rate constant, viz.,

$$k = k_r + k_{nr} + k_p + k_q[\mathrm{Q}]. \tag{2.2}$$

Thus, the fluorescence follows first-order kinetics and its intensity decays exponentially in time. The fluorescence lifetime, τ, is the time required for the fluorescence intensity to fall to 1/e of its initial value following pulsed excitation of short duration. The relationship of lifetime to the rate constant is a reciprocal one, i.e.,

$$\tau = {}^{1}\!/_{k}. \tag{2.3}$$

Fluorescence lifetimes are generally in the range of 10^{-9} s - 10^{-7} s.[9] Lifetimes are affected by a variety of factors, including solvent, other solutes, and dissolved oxygen. There are two standard methods in use today for measuring fluorescence lifetimes: pulsed or time-resolved techniques and harmonic or phase-modulated techniques.[25,26]

2.2 Lifetime Measurement Techniques

The nature of excited state lifetime determinations and the pitfalls associated with the different techniques are discussed in numerous sources, including a monograph by Demas.[25] The two principal approaches, pulsed and phase-modulated techniques, are discussed in this section, with emphasis on the former since it was the one used in this research.

It is convenient to relate lifetime determination to signal processing theory since fluorescence decay is, in reality, a time domain signal. Consider a fluorescing species as an analog system, like an electronic black box. The excitation photons can be viewed as simply an input, $e(t)$, to the system. The output of the system is the fluorescence. $f(t)$, i.e.,

$$f(t) = e(t) * y(t)\,, \tag{2.4}$$

where $*$ denotes the mathematical operation of convolution, and $y(t)$ is the impulse response of the fluorescence. Because these time-dependent signals are measured with a detector and electronic signal processors that impose their own time dependence, additional terms must be incorporated into the equation. The system now consists of the solution containing the species of interest: the detector, whose impulse response is $d(t)$; and processing electronics, whose impulse response is $l(t)$. The instrumentally observed fluorescence is

$$f(t) = e(t) * y(t) * d(t) * l(t)\,. \tag{2.5}$$

Since the fundamentally interesting term here is the impulse response of the species decay, i.e., $y(t)$, terms in (2.5) can be combined such that

$$f(t) = E(t) * y(t), \tag{2.6}$$

with

$$E(t) = e(t) * d(t) * l(t). \tag{2.7}$$

$E(t)$ is called the instrument response function, and it can be defined as the input to the system, $y(t)$, whose output is $f(t)$. The convolution in (2.5) is defined by

$$f(t) = \int_{-\infty}^{\infty} E(x) y(t - x) dx = \int_{-\infty}^{\infty} E(t - x) y(x) dx. \tag{2.8}$$

Since the fluorescence derives from an actual physical process, the system is causal, i.e., the signal at any time does not depend on future values of the input to the system.[27] In fact, since the impulse response is a physical (therefore, causal) system and $y(t)$ is a causal system, the result is

$$f(t) = \int_0^t E(x) y(t - x) dx, \tag{2.9}$$

for all $t > 0$ and zero for $t < 0$. Since $y(t)$ has the form

$$y(t) = e^{-kt}, \tag{2.10}$$

for all $t \geq 0$ and zero for $t < 0$, (2.8) becomes

$$f(t) = e^{-kt} \int_0^t E(x) e^{kx} dx, \tag{2.11}$$

for all $t > 0$ and zero for $t < 0$, which is the common form for the convolution of an excitation with an exponential decay.

It is important to note that $E(t)$ is not determined by just the excitation pulse shape itself. Rather, it reflects the response of the entire system except for the response of the fluorophore itself. The formulation of the relationship in (2.10) is important when considering how to extract the fluorescence lifetime from the data provided by pulsed and phase-modulated techniques.

2.2.1 Pulsed Techniques

In the pulsed methods, one attempts to measure the lifetime directly by exciting a solution containing the species of interest with a short duration burst of photons. The emission intensity is observed as a function of time following the excitation. Ideally, the response time of the detector and signal processor are much shorter than the decay, in which case one observes the true decay. Complications of this simple process occur because of three factors: finite excitation pulse width, i.e., $e(t)$; the detection response time, i.e., $d(t)$ with $l(t)$; and experimental noise.

There are a variety of excitation sources one can use to generate intense short pulses of UV and visible light. Three of the more popular sources are (1) gas-filled discharge lamps, which operate at a few kilohertz (KHz) with pulse durations of a few nanoseconds; (2) low-repetition rate (~100 Hz) excimer, nitrogen, and Nd:YAG lasers with typical pulse durations of a few nanoseconds; and (3) mode-locked and cavity dumped lasers with high-repetition rates (up to 80 MHz) and pulse durations in the picosecond range.[28]

Of course, the selection of excitation source may depend upon the fluorophore under investigation, but it is also affected by the choice of detection scheme. Ideally, the pulse duration is chosen to be much shorter than the lifetime, in which case one could consider

$e(t)$ a Dirac delta function, $\delta(t)$. If this is so, then one need only consider the response of the detection system in extracting the lifetime. However, if the pulse duration is not a delta function, the data will be *smeared*, and deconvolution is necessary.

Pulsed detection schemes commonly used in measuring lifetimes in the nanosecond and subnanosecond regime include (1) time-correlated single-photon counting (TCSPC) which can measure picosecond lifetimes, (2) streak cameras which can also measure picosecond lifetimes, and (3) oscilloscopes.

TCSPC employs two PMTs, signal amplifiers, constant fraction discriminators set up in parallel, a time-to-amplitude converter (TAC), and a multichannel analyzer (MCA).[29] The excitation pulse simultaneously illuminates the sample and the triggering PMT. The triggering PMT detects the excitation pulse and sends an amplified "start signal" to the TAC to start the ramp generator of the TAC. If a fluorescence *event* is detected by the second PMT, a stop pulse is generated, and the TAC converts the ramp value to a time. The TAC is programmed to stop at a predetermined value if no event occurs. If an event occurs, the MCA places a count into the appropriate memory channel, thus building a histogram of events as a function of time. If multiple photon events reach the second PMT, the TAC will only see the first one and miss subsequent ones. To prevent this so-called "photon pileup," an average count rate of 0.01 to 0.05 per pulse is desirable.[30]

TCSPC has many advantages. It is one of the most sensitive lifetime approaches owing to its need for low intensity fluorescence. It can measure decays much shorter than its system response would indicate. This results from the triggering and counting mechanisms. The start and stop signals are taken from the leading edge of the excitation pulse and fluorescence, respectively, as detected by the PMT and determined by the CFD. Thus, a PMT with a several nanosecond response time may be used to measure a nanosecond lifetime directly (provided, of course, the excitation source is a $\delta(t)$). Since TCSPC is a counting experiment, it follows well-defined Poisson statistics, so error analysis is greatly simplified.[25] TCSPC also has a large linear dynamic range, i.e., it has the ability to measure a wide (time) range of decay processes.

The major disadvantage in TCSPC is the time required to collect a profile. For good statistics, the recommended minimum of counts in the peak channel is 100,000.[30] With low repetition rate sources, acquisition of a decay curve can take many minutes or even hours. Mode-locked lasers are favored for TCSPC owing to their high repetition rates.

Streak cameras can record very fast processes in real time by converting time domain information into spatial information. The process is as follows: (1) Incident photons are converted by the photoelectric effect to electrons at a photocathode. (2) The photoelectrons are accelerated by a potential difference into an electron deflection field, which is varied linearly in time. (3) The photoelectron beam is swept across a microchannel plate, and individual electrons are amplified within the microchannels before striking a phosphor screen. (4) The image from the phosphor is recorded by a vidicon or CCD. Streak cameras provide very high temporal resolution, having been extended into the femtosecond range.[29] However, it is difficult to get both high temporal resolution (<10 ps) and long observation times (>5 ns) simultaneously. Streak cameras are also very expensive.

Oscilloscopes are, in general, the most inexpensive of the methods discussed here to measure lifetimes. In most oscilloscope-based lifetime studies, the anode signal from a

PMT is directly connected to an input channel of the scope. Oscilloscopes come in two basic varieties, analog and digital. Analog oscilloscopes are rarely used anymore for short lifetime (nanosecond regime) measurements, owing to the tediousness of the data processing. Formerly, images of the analog oscilloscope were captured photographically and the decay curves digitized manually for subsequent mathematical processing. Digital oscilloscopes or digital storage oscilloscopes (DSO) represent the state of the art in oscilloscope technology, and many are capable of recording transient signals in real-time or averaging repetitive transients.

DSOs acquire analog signals, the PMT output, by sampling and digitizing. As the analog input enters the DSO, the analog-to-digital converter (ADC) selects or samples the value of the signal, a voltage, at discrete points in time. The voltage is then digitized and displayed and/or recorded to be used later. This acquisition process generates a quantized signal, which is a mapping from a continuous domain space to a discrete range space. By acquiring many observations of the same transient, one can improve the S/N.

Several properties of the DSO affect the quality of the data it can produce. The ability of a DSO to accurately measure short duration signals is determined by the maximum sampling rate and the maximum analog bandwidth. The maximum sampling rate indicates the number of samples per second the ADC can acquire. The larger the sampling rate, the greater the time resolution the DSO can display. The rise time of the DSO is a measure of the DSOs ability to respond to rapid changes in signal. It is defined by the following formula:

$$t_r = \frac{400}{f_a}, \tag{2.12}$$

where t_r is the rise time in nanoseconds and f_a is the analog bandwidth in megahertz, which specifies the frequency range the DSO can accurately measure.[31] Rise times on good DSOs with sampling rates of 2 GS/s can be as short as 800 ps.

2.2.2 Phase-modulated Techniques

Phase-modulated spectroscopy is a method of indirectly measuring fluorescence lifetime by rapidly modulating the excitation light while simultaneously modulating the high voltage to the detector at almost the same rate.[32] The phase of the emission modulator is shifted relative to the excitation to extract the phase dependence of the fluorescence intensity. With commercially available equipment capable of modulation frequencies of 250 MHz, one can measure lifetimes as short as 1 ps.

The theory and practice of phase-modulated spectroscopy has been discussed in the literature.[25,26] The time-dependent fluorescence signal, $D(t)$, generated in a phase-modulated instrument with angular modulation frequency, ω (equal to $2\pi f$, where f is the modulation frequency in hertz), and phase-shift angle, ϕ, is given by

$$D(t) = A'\left(1 + m_{ex} m \sin(\omega t - \phi)\right), \qquad (2.13)$$

where A' is the wavelength-dependent, steady-state (DC) component of the fluorescence emission, m_{ex} is the modulation depth of the excitation (i.e., the ratio of the AC amplitude to the DC intensity), and m is the demodulation factor (given by the modulation depth of the fluorescence divided by m_{ex}).

The fluorescence lifetime may be calculated from the phase shift, τ_p, or the demodulation, τ_m, using

$$\tau_p = \frac{\tan\phi}{\omega}, \tag{2-14}$$

or

$$\tau_m = \frac{1}{\omega}\left(\frac{1}{m^2} - 1\right)^{\frac{1}{2}}, \tag{2-15}$$

respectively.

The phase-modulated method provides a fast, precise, and accurate way to measure fluorescence lifetimes in the laboratory. Unfortunately, when considering UV-visible fluorescence measurements, the phase-modulated method has lower S/N than the pulsed methods. While the state-of-the-art is advancing rapidly, in its present stage of development, phase-modulated instrumentation is more expensive and less amenable to fiber-optic and field experiments than the DSO-PMT method used in this work.

3. FACTOR ANALYSIS

Malinowski[33] has defined factor analysis as “a multivariate technique for reducing matrices of data to their lowest dimensionality by the use of orthogonal factor space and transformations that yield predictions and/or recognizable factors.” Some might consider the phrase “use of orthogonal factor space” overly restrictive, since many factor analysis methods employ non-orthogonal factors; but this definition provides an excellent starting point. Factor analysis methods for fluorescence data can be roughly separated into two areas: calibration (and prediction) and, what will be called here, profile extraction.

The aim of calibration is to extract quantitative information (e.g., concentrations) about the constituents of an unknown sample. Calibration is usually performed as a two-step process. First, in the calibration step, a concentration-response model is constructed from the instrumental response to standards. In the following prediction step, the instrumental response of an unknown is measured, and the concentration of the unknown is estimated, using the model from the calibration step.

The objective of profile extraction is qualitative information, in the form of recognizable factors, about the constituents of an unknown sample. These recognizable factors would be in the form of pure component spectra, time decay profiles, etc., which Malinowski[33] would identify as *real factors*. For example, from an EEM of a two-component solution containing benzene and naphthalene, one would attempt to extract the benzene and naphthalene excitation and emission spectra. Profile extraction can be performed with or without standards. As Malinowski’s definition suggests, there is sometimes overlap between the processes of calibration and profile extraction.

For environmental applications, one may have either or both goals in mind. The concern might be what toxic materials are present and at what concentrations in a soil or groundwater sample. To accomplish the quantitative goal using fluorescence data, one must develop a calibration model based on the relationship between fluorescence intensity and concentration. The qualitative goal is accomplished by extracting predictions of spectra or fluorescence decay curves and comparing them with entries in a database to identify the compound(s). For either goal, a sound understanding of the physical and mathematical nature of the fluorescence data is an important factor in constructing effective factor analysis schemes.

The mathematical nature of different types of fluorescence data is addressed in this chapter. Methods of calibration and profile extraction based on the type of data collected are also presented.

3.1 Fluorescence Data as Tensors

Tensorial notation was first introduced to describe multivariate calibration in the field of chemometrics by Sanchez and Kowalski.[34,35] Since then, Kowalski et al.[5,7,36-38] have disseminated tensor terminology, borrowed from mathematics and physics, to categorize the types of data acquired by various instruments used by chemists. Objections to the use of the term tensor in chemometrics have been raised because there are some tensor conventions which do not carry over into chemometrics.[39] However, there appear to be enough proponents to carry it into more general use. Owing to Kowalski's influence in this area of chemometrics, much of what follows is modeled after papers he has published on this subject.[5]

Tensor order in chemometrics indicates the number of modes inherent in a data set. Tensor order notation follows the same pattern as array order used in this thesis (see Definition A.22). The notation can be further extended to the instrumentation that collects the data to designate the tensor order of an instrument.

The order of an instrument provides insight into the flexibility one has available in the data analysis and the type of information to be gleaned from the data produced by such an instrument. The order of an instrument indicates the maximum tensor order of data that can be obtained on a single sample of analyte. Here, the focus will be on fluorescence instruments and the various orders of data one can collect with them.

3.1.1 Zeroth-order Instrumentation

Measurement of fluorescence intensity of a single luminescent species in solution at a single excitation wavelength and a single emission wavelength yields a scalar datum, f, which is a zeroth-order tensor. An instrument which can only measure zeroth-order data is appropriately called a zeroth-order instrument. An example of a zeroth-order fluorescence instrument would be one with a mercury penlamp excitation source and a combination optical filter and PMT detector. The concentration of the luminescent species present is proportional to f. Therefore, f is a quantitative datum, but it provides no qualitative information concerning the luminescent species. In fact, the qualitative information, viz., that only a single luminescent species is present, would have to be known before the fluorescence measurement. Zeroth-order data are thus very limited in terms of the information they provide to the analyst.

3.1.2 First-order Instrumentation

Measurement of fluorescence intensity as a function of emission wavelength, $f(\lambda)$, is a concept familiar to chemists. A decade ago, data acquisition with continuous scanning monochromators and strip chart recorders was common. The practice of most spectroscopists today is to measure spectra at discrete wavelengths with electronic recording of the data. Borrowing a term from signal processing, we refer to this as a quantized signal. An ordered collection of these quantized signals, $f(\lambda_i)$, constitutes a vector, **f**, directed from the origin in *n*-dimensional space (*n*-space), i.e.,

$$\mathbf{f} = \begin{bmatrix} f(\lambda_1) \\ f(\lambda_2) \\ \vdots \\ f(\lambda_n) \end{bmatrix}, \tag{3.1}$$

where the subscript denotes the *i*th wavelength at which an intensity was measured. A quantized spectrum represents first-order data, and an instrument capable of measuring at best such a spectrum is a first-order instrument.

An excitation spectrum, which is obtained by recording the fluorescence intensity at a fixed emission wavelength while the excitation wavelength is varied, also represents first-order data. Fluorescence decay data, $f(t)$, are also first-order data and are measured by a first-order instrument.

Two important properties of a vector are its length and its direction. When a single luminescent species is present, the length or norm of the spectral vector, $\|\mathbf{f}\|$, provides a measure of the amount of emitter present. A valuable attribute of the vector over zeroth-order data is that it provides many estimates of the intensity in a single experiment.

First-order data therefore provide a form of built-in signal averaging. The other main property of the spectral vector, its direction, gives insight into the identity of the emitter; thus, with first-order data, one gets the benefit of qualitative information.

First-order fluorescence data also include the emission intensity as a function of (1) excitation wavelength, (2) fluorescence decay time, or (3) polarization of the emitted photons relative to the excitation photons.

3.1.3 Second-order Instrumentation

One can generate second-order data by combining first-order measurement modes. For example, as mentioned in the introduction, an EEM is obtained by scanning the emission wavelength at several different excitation wavelengths. Similarly, one could vary wavelength and time to form a WTM. Such data can provide qualitative and quantitative information like first-order data, but with the additional benefits described in Section 3.2.3. Second-order fluorescence instruments capable of measuring EEMs have been in use for some time. Commercial versions of these instruments are now common. Instruments for measuring WTMs on a nanosecond or shorter time scale are a more recent innovation and are not commercially available at this time.

3.1.4 Third-order Instrumentation

A three-mode array (3-array) of data called a time-resolved excitation-emission matrix (TREEM) combines the modes of excitation wavelength, emission wavelength, and time domain. Instruments which generate these data are usually "souped-up" versions of the second-order instruments which produce WTMs and WFMs, whose third-order descendant is the excitation-emission frequency array (EEFA). If time domain data are available, the modification to scan both excitation and emission wavelengths is rather trivial. The ROST™ is a third-order instrument since it can collect TREEMs.

The advantages of this type of data will be discussed in Section 3.3.2.

3.1.5 Fourth-order Instrumentation

Optical spectroscopic instruments designed to obtain fourth-order data are possible, but are of less obvious usefulness. A time-resolved excitation-emission-anisotropy instrument may be of interest to biochemists[40] interested in the properties of fluorophores in living cells and tissue.

3.2 Tensorial Calibration

One seeks by calibration to establish a mathematical and statistical model of the relationship between experimental data and analyte concentration to estimate concentrations in an unknown sample. The usual procedure is to create a set of standards, with known concentrations, and measure the spectral response. The model is then created; and, in a prediction step, the concentration of the unknown is estimated. A linear relationship between the signal from the instrument and the desired parameter is desirable since it simplifies data analysis. The ability to determine concentrations of multiple analytes in mixtures and performance of the calibration model in the presence of interferences is influenced by the tensor order of the data.

An *n*th-order instrument generates *n*th-order data by definition. Multiple samples of calibration data (standards) can be assembled into an $(n + 1)$-order tensor. A calibration set of individually zeroth-order data, i.e., scalars, is a first-order tensor, i.e., vector; a set of first-order calibration data set represents a second-order tensor; etc. However, note that the order of calibration follows from the order of the instrument and not from the order of the calibration data set.

3.2.1 Zeroth-order Calibration

Zeroth-order calibration is a very common feature in data analysis. The methodology is simple, the analysis is represented very well graphically, and it can easily be performed with a hand calculator. The statistics are also simple, well-defined, and easy to calculate. Unfortunately, there is also very little flexibility with zeroth-order calibration. One cannot use zeroth-order data to simultaneously analyze multiple analytes in a mixture. Any interference beyond a constant offset in the data distorts the model.[5,6] Interferences present in the unknown sample, but not represented in the calibration set, cannot be detected in the prediction and will bias the parameter (concentration) estimate. A measurement which is highly selective for the target analyte is the most favorable case for zeroth-order calibration.

3.2.2 First-order Calibration

First-order calibration is both more powerful and more complex than zeroth-order calibration. Fortunately, there are several analysis algorithms whose mathematical and statistical properties have been extensively investigated. Among these are classical multivariate least squares (MLS), inverse least squares (ILS), principal component regression (PCR), and partial least squares (PLS), with PLS, PCR, and MLS being the most popular. First-order data allows for analysis of multicomponent mixtures with a maximum number of analytes equal to the number of elements in the spectral vector. Interferences can also be accommodated in the calibration model when using ILS, PLS, and PCR.

ILS, PCR, and PLS are actually all inverse least squares techniques,[5-7,41,42] in that they model concentration as a function of instrument response, opposite to the way analytical expressions are usually presented. For example, Beer's Law is often written

$$A = \varepsilon\ lc\,, \tag{3.2}$$

where A is the absorbance, ε is the molar absorptivity or extinction coefficient, l is the path length of the solution, and c is the concentration. The inverse relationship may be written as

$$c = bA\,, \tag{3.3}$$

where b is a model parameter which defines the inverse relationship. The advantage of this methodology is that the analysis is invariant with respect to the number of analytes in the model. The concentration of any spectrally active species may be determined independently of the others provided their spectral responses are modeled. In fact, in the method referred to as PLS1,[42] each analyte in a multicomponent mixture is modeled separately, and predictions on the unknown mixture are conducted independently in turn. The analysis for each component treats the mixture as if it contains only one analyte and numerous interferences. The limitation in this case is that each interference present in the unknown must be included in the model or the prediction will be biased. However, the presence of interferences in the unknown but not in the model can be detected by comparing the unknown spectrum to the standard spectra.

Data pretreatment is another area given much attention in the literature. Methods such as mean centering and variance scaling of the data are the two most often discussed. Mean centering is accomplished by subtracting the average value of the observations of a variable from each observation for that variable, i.e.,

$$f_{ij}(centered) = f_{ij} - \bar{f}_j \tag{3.4}$$

where f_{ij} is the fluorescence intensity of the *i*th sample at the *j*th wavelength and $\bar{f}_j$ is the mean of the fluorescence intensities at the *j*th wavelength, i.e.,

$$\bar{f}_j = \frac{\sum_i^N f_{ij}}{N}. \tag{3.5}$$

Mean centering accommodates data with a non-zero baseline. Scaling to unit variance is accomplished by

$$f_{ij}(scaled) = \frac{f_{ij}}{s_j} \tag{3.6}$$

where s_j is the standard deviation of the *j*th wavelength, i.e.,

$$s_j = \left(\frac{\sum_i^N \left(f_{ij} - \bar{f}_j \right)^2}{N-1} \right)^{1/2}. \tag{3.7}$$

Performing both mean centering and scaling to unit variance is referred to as standardization in statistical texts and autoscaling to unit variance in chemometrics literature. Standardizing data is usually performed on data which are measured on different scales with large range differences or if the units are not compatible.

With the appropriate data pretreatment and the optimum algorithm, first-order calibration can be very powerful. But, as one might expect, even higher order data increase versatility.

3.2.3 Second-order Calibration

The simplest second-order data follow a bilinear model, i.e., the second-order data array for a pure component can be decomposed into a dyad. If the data conform to the bilinear model, calibration yields a third-order tensor in the calibration model which is trilinear. Trilinear data are highly desirable since they allow for the following: (1) calibration with a single standard, (2) calibration and prediction in a single algorithm, (3) calibration and accurate prediction in the presence of interferences, (4) decomposition of the array to provide factor matrices containing the pure component spectra of the two spectral modes, and (5) the possibility of a unique decomposition of the array. This kind of second-order data is very powerful, especially in terms of the qualitative information that can be derived. Unfortunately, the mathematical analysis is complex, and the statistics are not well understood. The complexity is even greater if the second-order data are not bilinear.

In cases where pure component second-order data are not bilinear, the method of analysis will depend more on the chemical and physical nature of the system. Some systems are amenable to analysis with restricted Tucker models which *may*, under certain constraints, result in a unique decomposition and pure component spectral profiles. Calibration in other non-bilinear systems by methods such as non-bilinear rank annihilation (NBRA)[43] or residual bilinearization (RBL)[44,45] may provide accurate predictions, but the ability to obtain qualitative information is lost in the former method.

3.3 Profile Extraction

When employing a profile extraction technique, one attempts to obtain the underlying first-order profiles from the data. These profiles may be pure component emission or excitation spectra, time decay profiles, or relative concentrations in different samples. Profile extraction provides the qualitative information one can use to identify an unknown luminescent species in a solution.

This type of analysis requires data in the form of an N-array, with $N \geq 2$, formed by a collection of first-order data, e.g., first-order calibration data; or it could be second-order data from a single solution, e.g., an EEM. This type of analysis is usually labeled by the order of the data undergoing analysis rather than the order of the instrument, which makes sense because it conveys the most specific information concerning the capabilities of the analysis. Thus, profile extraction on an EEM, which is a 2-array, is called 2-mode profile extraction, and profile extraction on a TREEM is called 3-mode profile extraction.

Profile extraction is necessary only if there is more than one fluorescing species in the sample. In cases where there is only a single fluorescing species, the analysis is trivial. As in calibration, profile extraction capability increases with the order of the data.

3.3.1 Two-mode Profile Extraction

Two-mode profile extraction is the most commonly applied factor analysis method. Perhaps this is so because there are so many ways to construct first-order data and many of the techniques have been around for a long time. Malinowski[33] divides the methods of factor analysis into three categories: (1) abstract factor analysis (AFA), (2) target factor analysis (TFA), and (3) special methods.

AFA involves transformations of sets of abstract factor matrices to obtain real factors. Abstract factors are obtained by eigenanalysis of the data matrix. This process is often called principal component analysis (PCA) or principal factor analysis (PFA). PFA involves the decomposition of the data matrix, **D**, into two orthogonal matrices, **R** and **C**, often referred to as the scores and loadings matrices, respectively, with

$$\mathbf{D} = \mathbf{R}\mathbf{C}'. \tag{3.8}$$

Malinowski[33] describes a number of methods of decomposing **D**. One of the most popular ones is singular value decomposition (SVD), in which D is represented by

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}'. \tag{3.9}$$

The vectors in **U** are commonly called the left singular vectors and are the eigenvectors of $\mathbf{D}\mathbf{D}'$; those in **V** are the right singular vectors and are the eigenvectors of $\mathbf{D}'\mathbf{D}$; and **S** is the diagonal matrix of singular values, which are the square roots of the eigenvalues of $\mathbf{D}\mathbf{D}'$ (and $\mathbf{D}'\mathbf{D}$). Equations (3.8) and (3.9) are related by setting $\mathbf{R} = \mathbf{U}\mathbf{S}$ and $\mathbf{C} = \mathbf{V}$. The column vectors of **R** and **C** represent different amounts of *variance* explained by the decomposition.

The term variance in multivariate analysis describes the amount of information contained in a matrix. It stems from the common use of the covariance matrix in multivariate analysis. If **G** is the Gramian association matrix formed from **D**, the total variance of **D** is the trace of **G**, i.e., tr(**G**). The total variance of a matrix is equal to the sum of the eigenvalues, λ, of its decomposition, i.e.,

$$\mathrm{tr}(\mathbf{G}) = \sum \lambda . \tag{3.10}$$

The magnitude of the individual eigenvalues represent the fraction of variance described by the respective eigenvectors.

If those factors that account for the bulk of the variance are retained and the others are eliminated, the model in (3.8) can be reduced to abstract representations of the data resulting from only real signal, not from noise. This objective leads to attempts to determine the array rank (see Definition A.28) of the matrix. The term *pseudorank* is also used, since a data matrix with experimental noise will almost always be full rank in the strictest mathematical sense. There are a variety of methods to deduce the number of significant factors in a decomposition of this type.

The first two methods examine the error represented in the eigenvalues. Malinowski[33] describes a factor indicator function, that estimates which eigenvalues of the data originate from error (or from noise). If the eigenvalues, λ_n, are ordered from largest to smallest (λ_1 being the largest), the indicator value for the *n*th eigenvector is given by

$$\mathrm{IND}(n) = \frac{1}{(c-n)^2}\left(\frac{\sum_{j=n+1}^{c}\lambda_j}{r(c-n)}\right)^{1/2}, \qquad (3.11)$$

where c and r are the dimensions of the data matrix with $c < r$. The rank of the data matrix is the index n for which $\mathrm{IND}(n)$ is minimized.

Malinowski[46] also developed an *F*-test based on error represented in eigenvalues. The *F*-value is computed from

$$F(1,c-n) = \frac{\sum_{j=n+1}^{c}(r-j+1)(c-j+1)}{(r-n+1)(c-n+1)}\frac{\lambda_n}{\sum_{j=n+1}^{c}\lambda_j}. \qquad (3.12)$$

The eigenvalues are examined from smallest to largest. The index of the first eigenvalue whose $F(1,c-n)$ exceeds the tabular value of *F* at the desired significance level is the estimated rank of the matrix.

Other methods examine the eigenvectors. The first-lag autocorrelation coefficient $C(\mathbf{u})$, which has been used to separate signal eigenvectors from noise eigenvectors, can be calculated in the following way:

$$C(\mathbf{u}) = \frac{\sum_{j=1}^{c} u_j \, u_{1+j}}{\mathbf{u} \cdot \mathbf{u}} \tag{3.13}$$

where c is the dimension of the eigenvector, $\mathbf{u}$. Proponents[47] of this method assert that the eigenvectors corresponding to real signal will be smoother than those of the eigenvectors associated with noise. Thus, the signal eigenvectors possess larger values of $C(\mathbf{u})$. An autocorrelation coefficient value of 0.5 has been suggested as a cutoff between real and error eigenvectors.

After the number of factors has been determined, the selected factors are transformed with a transformation matrix, **T**. Transformation is straightforward.

$$\mathbf{X} = \mathbf{RT} \tag{3.14}$$

and

$$\mathbf{Y}' = \mathbf{T}^{-1}\mathbf{C}' \tag{3.15}$$

are the transformations of **R** into **X** and **C** into **Y**. **T** is a square matrix whose dimensions equal the number of selected factors. Its elements are usually determined by iterative transformations until the vectors **X** and **Y** meet some desired criterion. The two basic types of transformations are orthogonal and oblique. Orthogonal transformations preserve the angular dependence between the original vectors. Oblique transformations allow variation of the angles between the vectors.

Orthogonal transformations are primarily used to find clusters or patterns in the data at hand. An example would be trying to establish a relationship between chemical composition and physical properties of lake sediment. However, an orthogonal transformation would not allow one to extract physically meaningful spectra from an EEM.

Oblique transformations are another way to find clusters in data in order to identify correlated behavior. Based on minimization criteria alone, oblique rotations seldom reproduce identifiable spectra. TFA[33] is a method which employs oblique transformations, although the transformations matrix is determined in an different way. Also, a constraint included in the transformation algorithm can produce good results.

TFA attempts to extract meaningful factors one at a time by comparing transformations of the abstract factors to target factors, i.e., real factors one suspects are part of the data. In TFA, one generates a *transformation vector*, $\mathbf{t}$, by fitting the target vector, $\mathbf{x}$, to the abstract factor matrix in a least squares sense, i.e.,

$$\mathbf{t} = \mathbf{R}^{+}\mathbf{x}. \tag{3.16}$$

The *predicted vector*, $\hat{\mathbf{x}}$, is then obtained by

$$\hat{\mathbf{x}} = \mathbf{R}\mathbf{t}. \tag{3.17}$$

Equations (3.16) and (3.17) can be combined as

$$\hat{\mathbf{x}} = \mathbf{R}\mathbf{R}^{+}\mathbf{x} \tag{3.18}$$

to reveal that $\hat{\mathbf{x}}$ is a least squares projection of $\mathbf{x}$ into the space spanned by the data as represented by the abstract factors. The predicted vector is subjected to a statistical test (*F*-test) to determine whether it is a valid real factor. TFA requires that the experimenter have some idea of what may be in the sample and to have target vectors on hand. Of course, if one has a database, this should not be a problem, but for an unknown which may have spectra not in the database, this would pose a problem.

Special methods of factor analysis encompass a wide variety of techniques. The most common are classified as evolutionary factor analysis (EFA) methods, which include modeling and self-modeling algorithms. EFA methods often employ constraints based on known physical properties of the measured parameters. A common constraint for wavelength-mode spectroscopic data is nonnegativity.[48-52] Exponential decay is the usual constraint, or model, for time mode data in spectroscopy and kinetics.[53-55] Unimodality and nonnegativity have been used as constraints in chromatography. Modeling and self-modeling methods of EFA lend themselves readily to analysis of EEMs and WTMs.

Modeling methods can be used very effectively with fluorescence lifetime data. One simply uses the model in (2.11) to fit the time domain data. This technique was used by Knorr and Harris[53] to extract lifetime and emission spectra from WTMs of two-component solutions. Fluorescence decay following pulsed excitation was measured at a number of wavelengths. The WTM, **D**, is represented as a bilinear array,

$$\mathbf{D} = \mathbf{W}\mathbf{T}' , \tag{3.19}$$

where **W** is the matrix of wavelength mode factors and **T** is the matrix of time mode factors. The convoluted decays were fit in a Simplex search algorithm using a single lifetime parameter for each fluorescing species. Following each iteration of the simplex algorithm, Knorr and Harris obtained a least squares fit of the time mode vector to estimate the wavelength factor matrix, i.e.,

$$\mathbf{W} = \mathbf{D}\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1} . \tag{3.20}$$

Many practitioners[54,55] now call this technique *global analysis*. It is finding extensive use in biochemistry, especially kinetics. Often, users will search parameter space for parameters in the time mode and the wavelength mode while employing the Marquardt[56] search algorithm, although a faster scheme would employ the Marquardt method and the least squares method as Knorr and Harris did. This allows a faster search of parameter space for the nonlinear part (lifetime) and a fast solution to the linear part (wavelength). Of course, if both modes are built on nonlinear models, such a scheme would not be possible. One could also use nonnegativity constraints on the values of the wavelength mode to force a realistic solution.

Self-modeling methods impose constraints on the analysis without using any strict model for the behavior of the data. For example, constraining an oblique transformation of abstract factors to produce only positive values in the wavelength domain has been used to decompose fluorescence of several mixtures into recognizable spectra.[57,58]

Self-modeling curve resolution is a popular form of this type of analysis. Setting constraints such as nonnegativity for spectra and using an alternating least squares algorithm have been used with success.[48-52] As usual, the data matrix, **D**, is represented as the product of two factor matrices, **A** and **B**, i.e.,

$$\mathbf{D} = \mathbf{A}\mathbf{B}'. \tag{3.21}$$

With some starting value for **A**, **B** is estimated by least squares, i.e.,

$$\hat{\mathbf{B}}' = \mathbf{A}^{+}\mathbf{D}, \tag{3.22}$$

and any elements of $\hat{\mathbf{B}}$ that are negative are set to zero. Then, **A** is estimated from $\hat{\mathbf{B}}$ using

$$\hat{\mathbf{A}}' = \mathbf{B}^{+}\mathbf{D}', \tag{3.23}$$

and any elements of $\hat{\mathbf{A}}$ that are negative are set to zero. The process is continued until a termination criterion, such as minimum residual (see Appendix B), is reached. One could also employ a nonnegative least squares (NNLS) algorithm.

EFA procedures are simple to understand, and they do not require prior knowledge about the components in the system. By following a simple model, e.g., exponential decay or imposing constraints and iterating, one can generate the real factors in many cases. However, there is a danger that one could get a result that is not the correct outcome as a consequence of what is called the rotation problem.

The rotation (or transformation) problem is a direct result of (3.14) and (3.15). When operating on a 2-array, one can construct any transformation matrix and use it in (3.14) and (3.15). The result is that the decomposition of **D** is not unique. Even employing constraints such as nonnegativity does not ensure uniqueness.[10] However, decomposition of trilinear 3-arrays does provide a unique solution.

3.3.2 Three-mode Profile Extraction

Third-order data that follow the trilinear model can be decomposed uniquely. Some of the algorithms for performing the decomposition employ eigenanalysis based procedures (EBP), while others use the parallel factors (PARAFAC) analysis. PARAFAC is also called alternating least squares (ALS) or, as we shall henceforth refer to it, three-mode alternating least squares (3M-ALS).

The formal treatment of EBPs can be found in Appendix C. They are used to obtain a direct solution to the trilinear decomposition. There are variants of the EBP, but these deal mostly with the method of generating the data slices to be used in subsequent eigenanalysis. These methods are fast and produce good results on synthetic data and on real data with small numbers of components and high S/N. This methodology has been applied to several types of spectroscopic data.[10,36,39,59-62] However, studies have shown that the presence of noise can be problematic and that EBPs followed by 3M-ALS is necessary.[10,63]

3M-ALS is an iterative approach to trilinear analysis. Its methodology is described in Appendix B. 3M-ALS usually produces good results even in the presence of noise. Unfortunately, the iterative approach can be very time consuming, requiring tens of thousands of iterations. In addition, there are some mathematical curiosities associated with 3M-ALS.

Kruskal et al.[64] reported a phenomenon they referred to as two-factor degeneracy. A two-factor degeneracy exists when two factors in a trilinear decomposition (1) are highly correlated in all three modes, (2) have very large and almost equal magnitudes, (3) have the overall effect of cancellation of one factor's contribution by the other.

Another phenomenon that has been associated with 3M-ALS is multiple local optima (MLO).[10] MLO appear when two starting sets of parameters yield different solutions to the 3M-ALS sequence. Mitchell[10] has suggested that MLO are a function of termination criteria in many cases.

One way to expedite the convergence of the 3M-ALS algorithm with real spectral data would be to combine it with some of the techniques used in self-modeling curve resolution. By imposing a nonnegativity constraint, the algorithm may be forced to converge sooner. Three-mode nonnegative alternating least squares (3M-NNALS) may also help to resolve factors difficult to separate with 3M-ALS. Such an algorithm can easily be created by setting all negative elements in each mode to zero after obtaining a solution for that mode.[65]

The danger is that the procedure may not converge to the unique solution if the data are noisy. Another problem with 3M-ALS is that there are few promising rank estimation procedures available.

It is ironic that while there are an abundance of rank estimation procedures for second-order data (specifically, 2-arrays), there is no mathematical formalism for determining the rank of a 3-array.[12] Thus, rank estimation is very difficult with trilinear data.

Often, TLD is performed on a 3-array for a number of trial ranks. The rank of trilinear data is determined by examining the resultant factors for each trial rank. When the true rank of the data is exceeded, resolutions should produce factors that model noise and will appear as noisy factors in all three modes.

4. PATTERN RECOGNITION

From the time people are very young, they are very adept at pattern recognition. A baby easily recognizes his mother's voice and face. The phenomenal capabilities of the brain allow human beings to detect extremely subtle differences in sounds, images, odors, and textures, to name a few.

For data analysis, it would certainly be beneficial to train a computer to recognize patterns and to place samples into different groups or classes. Unfortunately, this is a nontrivial task. What humans perform innately must be mathematically modeled for the computer to perform. Important characteristics about the items to be grouped must be identified and given a numerical value. These are called *descriptors* of the data. Also, the method one uses to group the data will depend on what information one has beforehand.

Pattern recognition methods fall into two categories: *supervised classification* and *unsupervised classification.*[66] Supervised classification, or simply classification, seeks to place items into one of two or more known groups. Unsupervised classification, or clustering, seeks to place items into groups based on similarities without association to previously established groups. To avoid confusion, henceforth, we will refer to supervised classification as classification and unsupervised classification as clustering.

4.1 Classification

Classification algorithms are developed based on specific rules for selecting membership in a class. While there are various rules for determining class membership, the ultimate goal is to place the unknown into the correct class. The quality of a classification rule is determined by the probability of misclassification. The occurrence of noise makes it unlikely that 100 percent classification accuracy can be obtained for experimental data. When the probability of misclassification is small, it is called an optimal classification rule.

4.1.1 Optimal Classification - Bayes' Rule

Several authors consider Bayes' rule to be *the* optimal classification rule.[67,68] Bayes' rule holds that one should place an item, whose data are in **x**, in the group with the highest posterior probability, $\mathrm{P}(G_i|\mathbf{x})$ (read as the probability of belonging to group G_i given the data in **x**.) Thus, one wants to assign the item to the group *i* such that

$$\mathrm{P}(G_i|\mathbf{x}) > \mathrm{P}(G_j|\mathbf{x}) \qquad \text{for all } j \neq i. \tag{4.1}$$

$\mathrm{P}(G_i|\mathbf{x})$ is not easily obtained in practice, though it may be realized through Bayes' theorem, viz.,

$$\mathrm{P}(G_i|\mathbf{x}) = \frac{\mathrm{P}(\mathbf{x}|G_i)\mathrm{P}(G_i)}{\sum_i \mathrm{P}(\mathbf{x}|G_i)\mathrm{P}(G_i)}, \tag{4.2}$$

where $\mathrm{P}(G_i)$ is the prior probability of occurrence, i.e., the probability of an item being a member of a group in a population (without regard to the descriptors used to classify it), $\mathrm{P}(\mathbf{x}|G_i)$ is the probability of acquiring the descriptors in **x** given the group G_i. Thus, the item with the largest value for $\mathrm{P}(\mathbf{x}|G_i)\mathrm{P}(G_i)$ will have the largest posterior probability, and the item would be assigned to the group which satisfies

$$\mathrm{P}(\mathbf{x}|G_i)\mathrm{P}(G_i) > \mathrm{P}(\mathbf{x}|G_j)\mathrm{P}(G_j) \qquad \text{for all } j \neq i. \tag{4.3}$$

Assume the descriptors in the groups are normally distributed with mean μ_i and covariance Σ_i. Then,

$$\mathrm{P}(\mathbf{x}|G_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)'\,\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right), \qquad (4.4)$$

where p is the number of descriptors in $\mathbf{x}$ and $|\Sigma_i|$ is the determinant of Σ_i. It can be shown that one can generate the *quadratic discriminant score,* d_i^Q, by combining the left side of (4.3) and (4.4) and eliminating constants that are the same for all populations,[69] i.e.,

$$d_i^Q = -\frac{1}{2}\ln|\Sigma_i| - \left(-\frac{1}{2}(\mathbf{x}-\mu_i)'\,\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right) + \ln \mathrm{P}(G_i). \qquad (4.5)$$

From (4.3), then, the item with descriptors, $\mathbf{x}$, would be allocated to the group with the largest quadratic discriminant score.

Further assume that the groups share a common covariance, Σ. It also can be shown that one can generate a *linear discriminant score* d_i analogous to (4.5).[69] However, a more convenient form for a linear discriminant is the Mahalanobis distance squared, δ_i^2,

$$\delta_i^2 = (\mathbf{x}-\mu_i)'\,\Sigma^{-1}(\mathbf{x}-\mu_i). \qquad (4.6)$$

Allocation of the item with $\mathbf{x}$ to group G_i is based on

$$-\frac{1}{2}\delta_i^2 + \ln\mathrm{P}(G_i) > -\frac{1}{2}\delta_j^2 + \ln\mathrm{P}(G_j) \quad \text{for all } j \neq i. \qquad (4.7)$$

Note that if the prior probabilities are equal, the inequality in (4.7) reduces to allocation based on the minimum δ_i^2.

In practice, the population covariance (for the case of common covariances), Σ, and means, μ_i, are not usually known. Instead, they are estimated from calibration sets of data giving the pooled covariance matrix, $\mathbf{S}_{\mathbf{pooled}}$, and the sample means, $\bar{\mathbf{x}}_i$, respectively. These are calculated by

$$\mathbf{S}_{\text{pooled}} = \frac{\sum_{i=1}^{L}(n_i - 1)\mathbf{S}_i}{\sum_{i=1}^{L}(n_i - 1)}, \tag{4.8}$$

where L is the number of groups, n_i is the number of items in the ith group. $\mathbf{S}_i$ is the sample covariance of the ith group, given by

$$\mathbf{S}_i = \frac{\sum_{k=1}^{n_i}(\mathbf{x}_{ki} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_i)'}{(n_i - 1)}, \tag{4.9}$$

and $\overline{\mathbf{x}}_i$ is the sample mean vector of the descriptors of the kth item in the ith group, given by

$$\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{x}_{ki} . \tag{4.10}$$

Optimal classification rules may include additional considerations.[69] One of these is the "cost" of misclassification, i.e., a number representing the potential value of a misclassification; e.g., an engineer classifying a failing bridge as safe would be more costly than the reverse.

4.1.2 Fisher's Method of Linear Discrimination

One objective of pattern recognition schemes is to find those aspects of the data which best describe the group. Feature selection (also referred to as feature extraction) endeavors to map the original descriptors into a lower dimensional space while providing greater group separation. Canonical discrimination or Fisher's linear discriminant function,[69] which is a feature selection technique for discriminating among several groups, is described in capsule form here.

Assume a common covariance matrix for the different groups and sample estimates rather than population statistics. Define the overall mean vector $\overline{\mathbf{x}}$ as

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{L}\sum_{k=1}^{n_i}\mathbf{x}_{ki}}{\sum_{i=1}^{L} n_i}. \tag{4.11}$$

Define the sample between groups matrix $\mathbf{B}_0$ as

$$\mathbf{B}_0 = \sum_{i}^{L}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'. \tag{4.12}$$

This is a measure of the spread of the means of the groups. Next, define the sample within groups matrix $\mathbf{W}$ as

$$\mathbf{W} = \sum_{i=1}^{L}\sum_{k=1}^{n_i}(\mathbf{x}_{ki} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_i)'. \tag{4.13}$$

$\mathbf{W}$ is a measure of the spread of the values within a group and is related to $\mathbf{S}_{\text{pooled}}$ by a constant.

The linear discriminant functions are determined by finding the eigenvectors $\mathbf{e}$ of $\mathbf{W}^{-1}\mathbf{B}_0$.[68,69] This optimizes $\mathrm{tr}(\mathbf{W}^{-1}\mathbf{B}_0)$, a measure of class separability.[68] This is accomplished by solving the eigenvalue equation,

$$\mathbf{W}^{-1}\mathbf{B}_0\mathbf{e} = \lambda\,\mathbf{e}. \tag{4.14}$$

Since $\mathbf{B}_0$ has rank, $L-1$, there will be only $r = min(p, L-1)$ nonzero eigenvalues, λ. Thus, only the r corresponding eigenvectors need to be utilized to generate features or linear discriminants without losing class separation. Following eigenanalysis, the eigenvectors are each scaled by

$$\mathbf{e}\,c = \mathbf{l}, \tag{4.15}$$

such that

$$\mathbf{l}\,\mathbf{S}_{\text{pooled}}\,\mathbf{l} = 1,$$

providing the coefficients **l** for a linear discriminant function.

The inner product of **l** and **x** produces the sample linear discriminant, *y*. Using the coefficients corresponding to the largest eigenvalue, one generates the sample first discriminant, $y_1 = \mathbf{l}'\mathbf{x}$. With the *p* by *r* matrix of the scaled significant eigenvectors **L**, one maps **x** into the new discriminant space, forming the sample discriminant vector $\mathbf{y} = \mathbf{L}'\mathbf{x}$.

If $r \leq 3$, the sample discriminants may be plotted against one another for each sample, showing the groups optimally separated in the discriminant space.

Using all of the *r* significant eigenvectors provides for a classification scheme identical to (4.6) and (4.7). The Mahalanobis distance squared, δ_i^2, is given by

$$\delta_i^2 = (\mathbf{y}-\bar{\mathbf{y}}_i)'(\mathbf{y}-\bar{\mathbf{y}}_i) = [\mathbf{L}'(\mathbf{x}-\bar{\mathbf{x}}_i)]'[\mathbf{L}'(\mathbf{x}-\bar{\mathbf{x}}_i)]. \tag{4.16}$$

4.2 Cluster Analysis

Cluster analysis is a method of separating items into groups based on where the descriptors of the data place them in space. As with all of the techniques discussed so far, there are a multitude of clustering techniques. One fast method which is relatively easy to use and understand and provides an easily interpreted graphic is agglomerative hierarchical cluster analysis (HCA).[69,70]

HCA operates by finding similarities (or dissimilarities depending on the measure one chooses to utilize) between items one at a time. It is an iterative process, taking *N* -1 steps for *N* items. Groups are formed as the items become linked together based on similarity. Each time a group is formed, it is treated as a single item. The measure of similarity at the time each group is formed becomes the level of that group in the hierarchy of groups. Thus, at the start of the process, there are *N* groups; and, at the end, there is only one, and the levels indicate the natural groupings. Each time a new group is formed, i.e., a linkage is formed, the similarities change to reflect the relationship of that group to the other groups.

Three common ways of forming linkages between groups are (1) single linkage, choosing the two items within the different groups with the greatest similarity between them and assigning that as the new measure; (2) complete linkage, choosing the two items within the different groups with the least similarity; (3) average linkage, creating an average over all the items between the groups. The choice of method will depend on the data; however, examination of all three methods is recommended. Different measures of similarity should also be explored.

The best measure of similarity will depend on the nature of the data under examination. A common measure of dissimilarity, perhaps useful as a first guide for all data, is distance. It is natural to associate objects based on how close they are in space. After all, that is the nature of groups. The distance between any two objects in coordinate space defined by their descriptor vectors, $d(\mathbf{x},\mathbf{y})$, can be determined using a variety of metrics. A versatile one is the Minkowski metric,[69,70] i.e.,

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{1/m}, \qquad (4.17)$$

where *p* is the number of descriptors in **x** and **y** and *m* defines the specific metric. For example, if $m = 1$, $d(\mathbf{x}, \mathbf{y})$ is the "city-block" or "Manhattan" distance, so called because it is analogous to measuring distance on a city grid while driving or walking. Using $m = 2$ results in the familiar Euclidean or "straight-line" distance. Relating each object to every other object is a simple task from the matrix of distances.

One measure of similarity which has been used for spectra is the angle cosine[70,71] (see Definition A.20). The angle cosine can take on values from -1 to 1, although with real nonnegative spectra, its values will only fall from 1 to 0. A high value of the angle cosine indicates more similarity: perfect agreement is indicated by the value 1. One notes a similarity to the familiar correlation coefficient, and the two are related. To make the similarity measure a minimum like distance , i.e., small value equals high similarity, the angle cosine is simply subtracted from one. The angle cosine has the added benefit that the magnitude of the individual vectors do not affect the correlation.

5. RESULTS AND DISCUSSION

The objective of this research was to examine chemometric methods that can be applied to determine the concentrations of luminescent species in solutions and mixtures and the identity of composite mixtures from fluorescence data obtained by the ROST™ or a similar instrument. Specifically, we investigated the chemometric analysis of time-resolved multimode fluorescence data. The background literature in this area is almost nonexistent, particularly for analysis of actual data, and we felt that investigation of a wide range of approaches would yield greater progress than in-depth studies of narrower topics. Our research should be viewed more as establishing a path for future research and validation of chemometric approaches than as a detailed comprehensive study.

We shall describe and discuss chemometric analysis of three data sets. The first two data sets were acquired for dilute solutions with only a few components. The first data set, Data Set One, was measured as a WTM-Concentration Array. The second set, Data Set Two, was measured as a TREEM. Our analysis goals in each case were to extract the spectral and temporal profiles for each luminescent species present in the solution.

Data Set Three consists of WTMs of four different fuel products present at three different concentrations on three different soil matrices. Our analysis goals for these data were to extract profiles for the luminescent species in the fuels, to classify the fuels using *a priori* information on the fuel types, and to perform cluster analysis to group the fuels without the benefit of *a priori* information.

5.1 Profile Extraction of Simple Systems

We defined a chemically simple system as one in which a small number of luminescent species are present at such sufficiently low concentrations that energy transfer effects are negligible. Such simple systems offer the best prospects for extracting chemically recognizable factors from the raw data. We made this assumption because simulated data utilized in most tests of chemometric methods resemble data from simple systems.[10,36,39]

5.1.1 Analysis of Three-mode Data With an EBP, 3M-ALS, and 3M-NNALS

We focus on several key aspects of trilinear analysis in this section. Trilinear three-mode data have the important inherent property that they can yield rotationally unique decompositions.[12] In the analysis by an EBP or 3M-ALS, one can either estimate the rank of the 3-array before analysis or perform the analysis for various assumed array ranks until a satisfactory decomposition is obtained; the latter is the usual procedure.

5.1.1.1 Data Set One: A WTM-Concentration Array

The first sample evaluated contained fluorene, pyrene, and naphthalene in aqueous solution. These compounds were selected because their spectra exhibit only modest overlap along the wavelength mode and their lifetimes are well-separated. Solutions were prepared by successive additions of aliquots of stock solutions of the analytes to 3 mL of water in a standard 1-cm internal path fused silica cuvette. The stock solutions were prepared in spectral grade methanol at concentrations of 1 ppm (± 7%), 100 ppm (± 5%), and 100 ppm (± 5%), for fluorene, pyrene, and naphthalene, respectively. A total of seven solutions (including the water blank) were prepared according to the design in Table 5.1.

Contour plots of the WTMs that were measured for each of the seven solutions are presented in Figure 5.1. Note that fiber optics were not used for light delivery and collection in these experiments. The excitation source was the fourth harmonic (266 nm) of a Nd:YAG laser (Spectra Physic GCR-12). The fluorescence emission was collected at right angles to the excitation and imaged with a lens onto the entrance slit of a 320 mm focal length monochromator (ISA HR-320). Decay profiles were quantized at 2 ns intervals over a 400 ns range; and emission spectral profiles were collected at 10 nm intervals over the wavelength range 280 nm - 480 nm. Therefore, each WTM consists of 201×21 or 4221 data elements.

Table 5.1. Experiment matrix for Data Set One

SOLUTION	Total Fluorene (ppb)	Total Pyrene (ppb)	Total Naphthalene (ppb)
1	0	0	0
2	0.33	0	0
3	0.33	33	0
4	0.33	33	33
5	1.0	33	33
6	1.0	67	33
7	1.0	67	100

5.1.1.1.1 Rank Estimation of Data Set One

With the knowledge that the solutions were prepared by standard additions, it is easy to deduce that there are at least four components represented in the data. The water Raman scatter is the dominant feature for WTM 1 (that of the blank) in Figure 5.1, and the additions of analyte are evident in WTMs 2, 3, and 4. The qualitative differences between WTM 4 and WTMs 5-7 are not nearly as easy to discern. A significant challenge for the

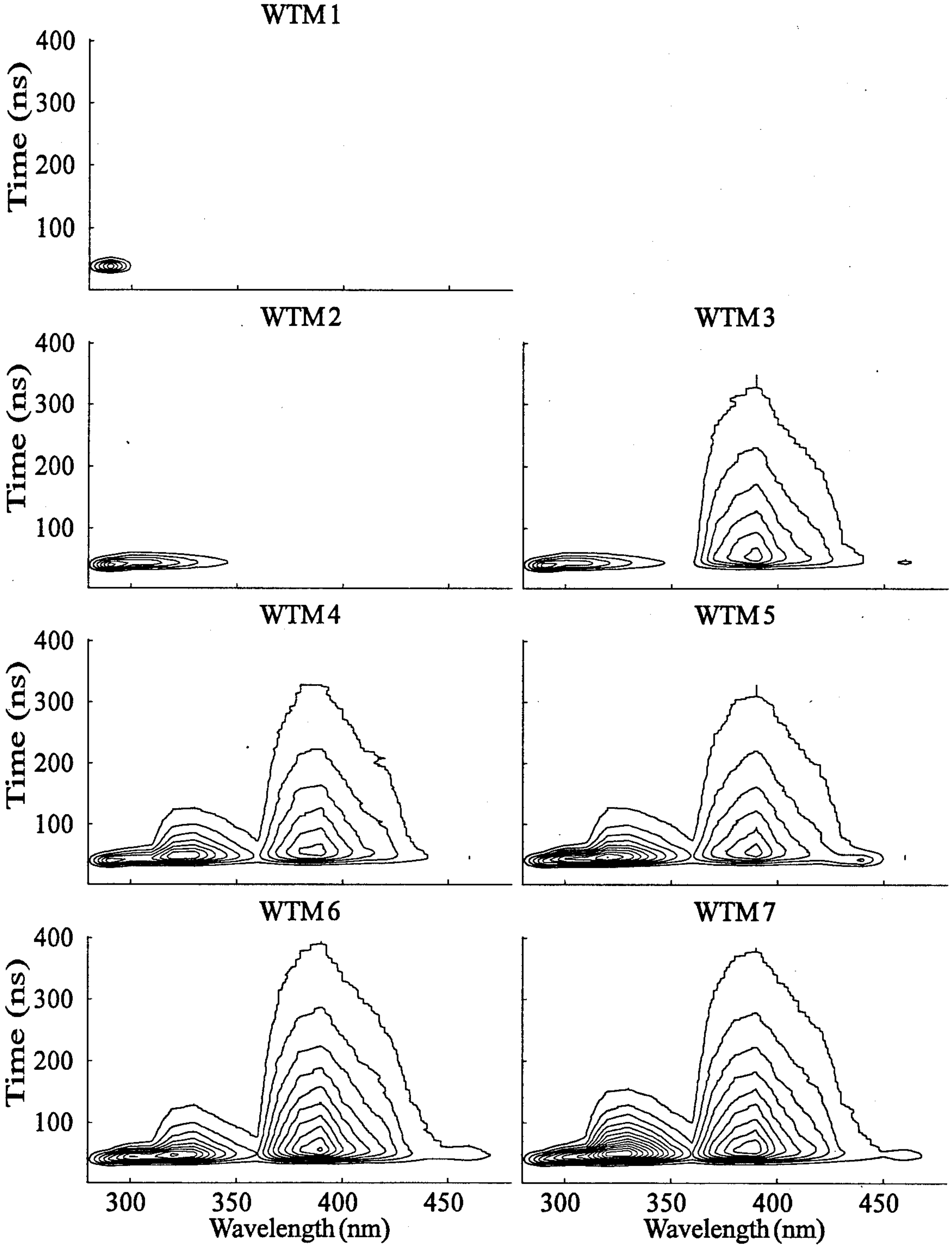


Figure 5.1. Contour plots of seven WTMs of Data Set One. Contours represent levels of equal fluorescence intensity. The weakest intensity is represented by the outermost contour in each WTM; it has a value of 0.025 (arbitrary units). Intensity increases by 0.025 for each successive contour.

chemometric methods, therefore, is to accurately predict the data ranks for individual WTMs and for the entire data set.

SVD were tested on this data set. Since the SVD can be performed for each individual WTM, the rank estimation methods were applied to each individual WTM. Malinowski's factor indicator function and *F*-test use only the eigenvalues, whereas the eigenvectors serve as input if the rank is estimated from the first-lag autocorrelation coefficient. Note that first-lag autocorrelation coefficients can be calculated in both the wavelength-mode space and the time-mode space. An autocorrelation value of 0.5, suggested for a cutoff between real and error eigenvectors, was employed here.

The solutions in Data Set One were constructed to contain one, two, or three added chemical components, and the water Raman signal represents a fourth component. The total number of components, which we refer to as the *a priori* (correct) data rank, is shown in Table 5.2 along with the rank estimations by the various methods. By inspection of Table 5.2, it is clear that the autocorrelation coefficient along the time mode is a totally unreliable rank predictor at the recommended 0.5 cutoff value. The agreement between *a priori* and estimated rank improves substantially when the cutoff value is increased to a higher value, e.g., +0.9; but at this stage, there is no strong justification for choosing a different cutoff value. The factor-indicator method and the *F*-test, which are based on eigenvalues, accurately predict the rank for WTMs 3 through 7, but overestimate the rank for WTMs 1 and 2. In fact, both methods actually give a higher rank for WTMs 1 and 2 than for the other WTMs even though the number of chemical components is definitely less. The wavelength-mode autocorrelation estimate exhibits reverse behavior. It agrees with the *a* *priori* rank for WTMs 1 and 2, but underestimates the rank by one for all of the solutions that contain four components. We speculated that the fact that the Raman scatter spectrum consists of a single non-zero intensity value at the wavelength interval used in the data collection is the reason.

Table 5.2. Rank estimates for Data Set One

SOLUTION	Rank[1] (with Raman)	IND[2]	*F*-test[3] (5 %)	$C(\mathbf{u}(\lambda))$[4]	$C(\mathbf{u}(t))$[4]
1	0(1)	5	4	1	7
2	1(2)	4	4	2	8
3	2(3)	3	3	2	17
4	3(4)	4	4	3	17
5	3(4)	4	4	3	18
6	3(4)	4	4	3	18
7	3(4)	4	4	3	17
Σ(WTMs)	3(4)	5	5	4	18

[1] Rank is equal to the number of emitting species; parenthetical value includes Raman.
[2] Malinowski's factor indicator.
[3] Malinowski's *F*-test for significance above the 5% level.
[4] Autocorrelation coefficient, values above +0.5.

Unfortunately, each of these methods, as applied normally, estimates the rank for a 2-array. This could cause a problem if a luminescent species is present in one solution, but not another; then the rank estimate of the 3-array would be deficient. For example, consider the estimate of the rank of a 3-array consisting of two WTMs where each WTM contained fluorescence from five luminescent species, but there were only four species common to both WTMs. In such a case, the rank of each WTM could be correctly estimated as four, but the rank of the 3-array would be underestimated. We have therefore tested the 2-array rank estimation methods to a compressed 3-array obtained by adding all of the individual WTMs. The results are contained in the last row of Table 5.2. Note that the method cannot provide a correct rank estimate if the rank of the 3-array is lower than the smaller dimension of the matrix. This is not a problem here since the smaller dimension of the new matrix obtained by summing the WTMs is 21.

This WTM summation procedure appears to cause the factor indicator function and the *F*-test to overestimate the rank by one. The wavelength-mode autocorrelation estimate is equal to the number of components in solution plus Raman scatter, but this may be fortuitous. Once again, the time-mode autocorrelation is poor using the +0.5 cutoff.

5.1.1.1.2 TLD of Data Set One

We next tested TLD of Data Set One for assumed ranks of two through six. Only selected results are shown here. Note that TLD treats the entire data set, which consists in this case of the seven WTMs shown in Figure 5.1. The first try at decomposition used the EBP of Sanchez and Kowalski,[36] and the rank four decomposition results are presented in Figure 5.2 along the emission wavelength, fluorescence decay time, and concentration modes. Factors Two and Four have negative intensity components in the wavelength and time modes, which are physically impossible; but the factors can nonetheless be reasonably assigned as follows: Factor One - naphthalene; Factor Two - fluorene; Factor Three - pyrene; Factor Four - water Raman scatter.

Next we tried 3M-ALS on these same data using the factors from the EBP as starting vectors. The convergence criterion for the 3M-ALS analysis sequence was a difference in consecutive norm of residuals less than 1×10^{-10} (this criterion was used for all calculations in this thesis). The algorithm provided a converged solution in 104 iterations. The factors for the rank four 3M-ALS decomposition are displayed in Figure 5.3. They improve upon

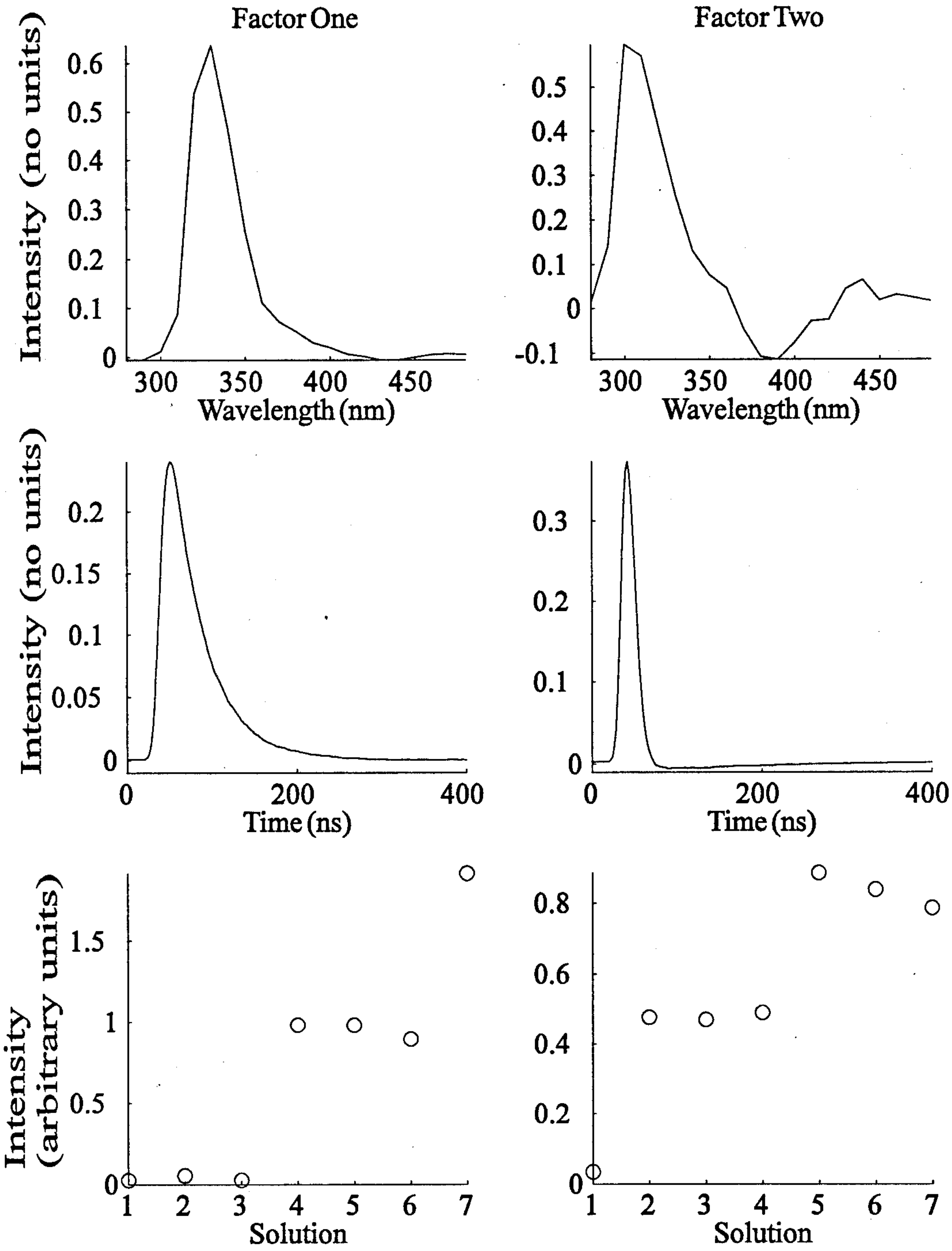


Figure 5.2. Rank 4 TLD of Data Set One using EBP of Sanchez and Kowalski.

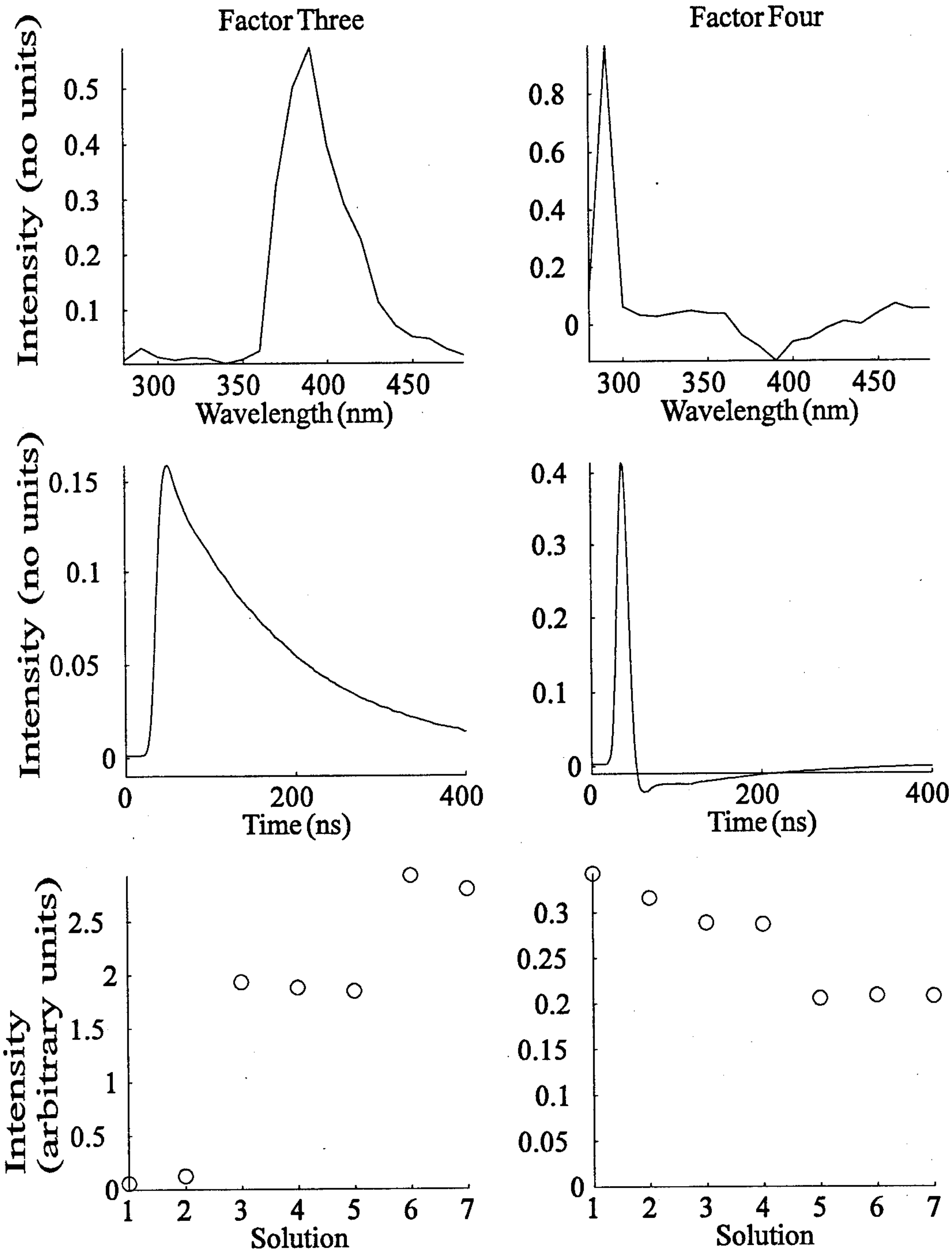


Figure 5.2. Continued.

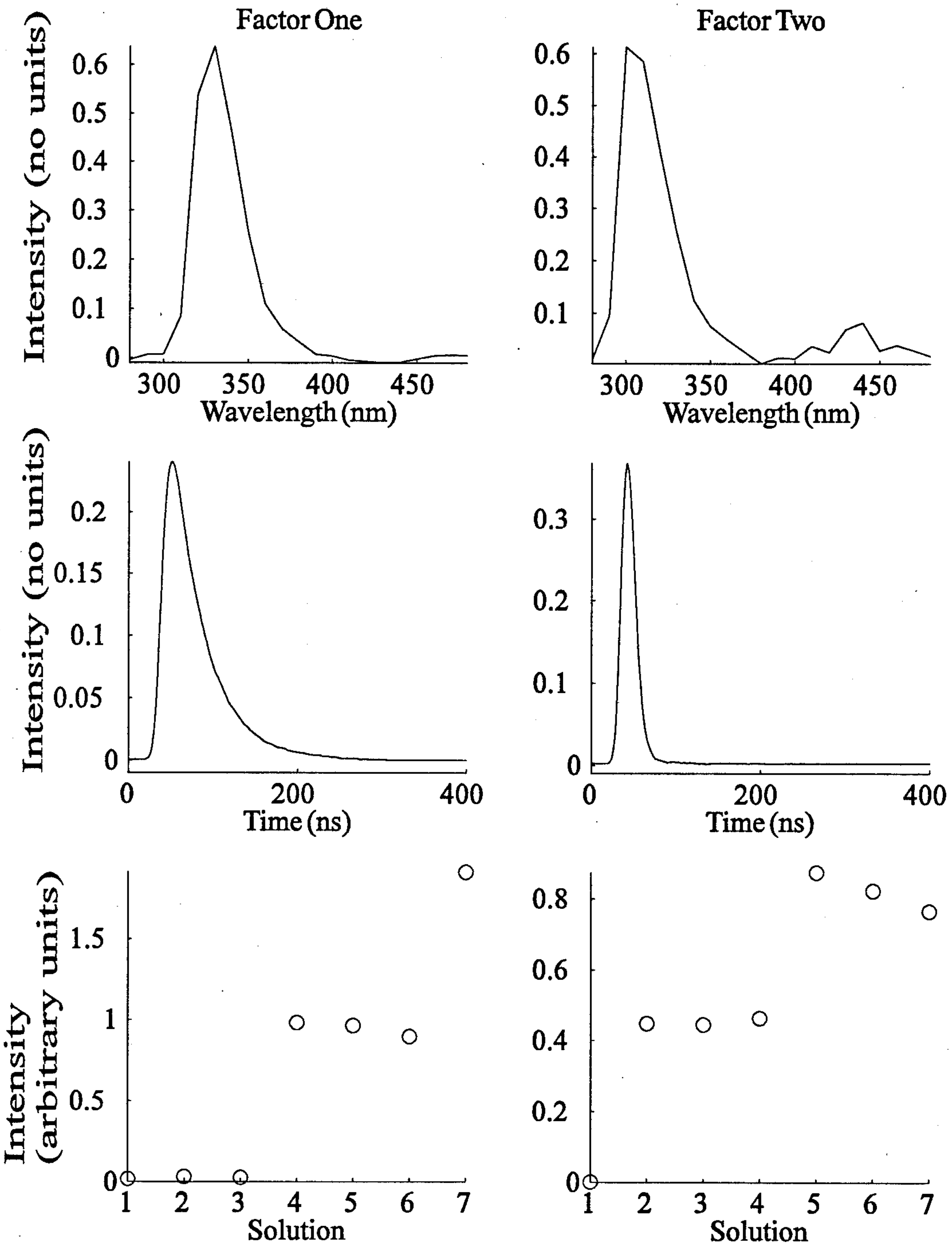


Figure 5.3. Rank 4 3M-ALS determined factors for Data Set One.

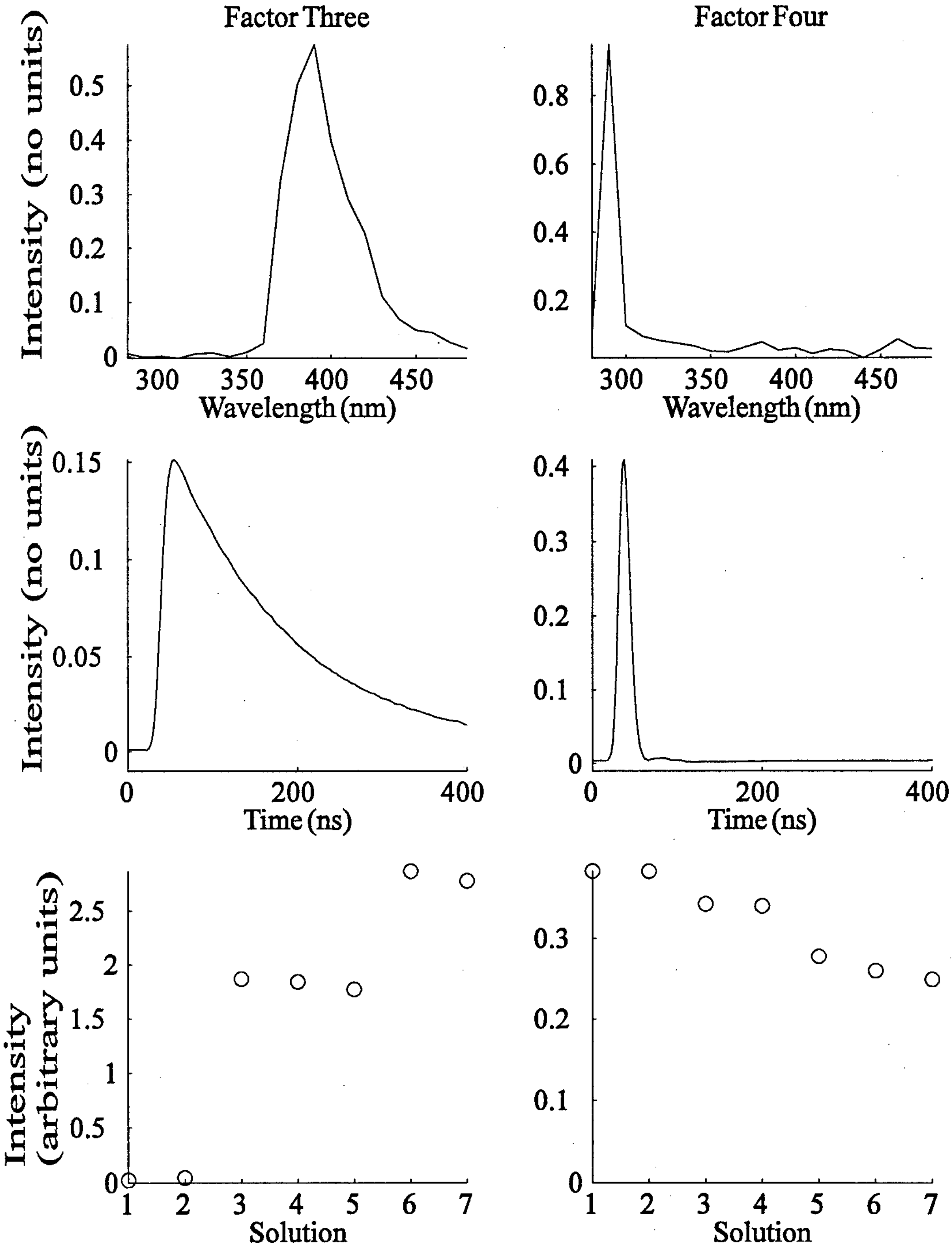


Figure 5.3. Continued.

the factors from the EBP since the negative-going features have been almost completely eliminated from the spectral and time-mode factors. The time mode of Factor Four (water Raman scatter) exhibits a slight bimodal character.

Three-mode NNALS (3M-NNALS) was also executed starting with EBP factors and with the same convergence criterion. The series converged after 93 iterations, a minor reduction compared with 3M-ALS. The results of this calculation are in Figure 5.4. Not surprisingly, the 3M-NNALS factors are practically indistinguishable from the factors presented in Figure 5.3.

Fluorescence lifetimes were computed with the phase plane method of Demas[25] from the time-mode profiles shown in Figure 5.4. The Raman scatter from WTM 1 was used as an estimate of instrument response function, $E(t)$. The results were 7 ns for fluorene, 40 ns for naphthalene, and 140 ns for pyrene, which compared favorably with independent determinations of 7 ns, 36 ns, and 127 ns, respectively, for these species in air-saturated water.[30] The fluorescence spectra also correlated well with the pure component spectra at this resolution.

The rank 5 decomposition provided an intriguing aspect to the analysis. A contaminant from an unknown source appeared as a fifth factor. The results of the EBP and the 3M-NNALS for the additional factor are displayed side-by-side in Figure 5.5. It is not clear at this time what the source of the fifth factor is since it only appears in one WTM, at approximately 440 nm in WTM 5 in Figure 5.1. The lifetime of this contaminant is very short, approximately 2 ns.

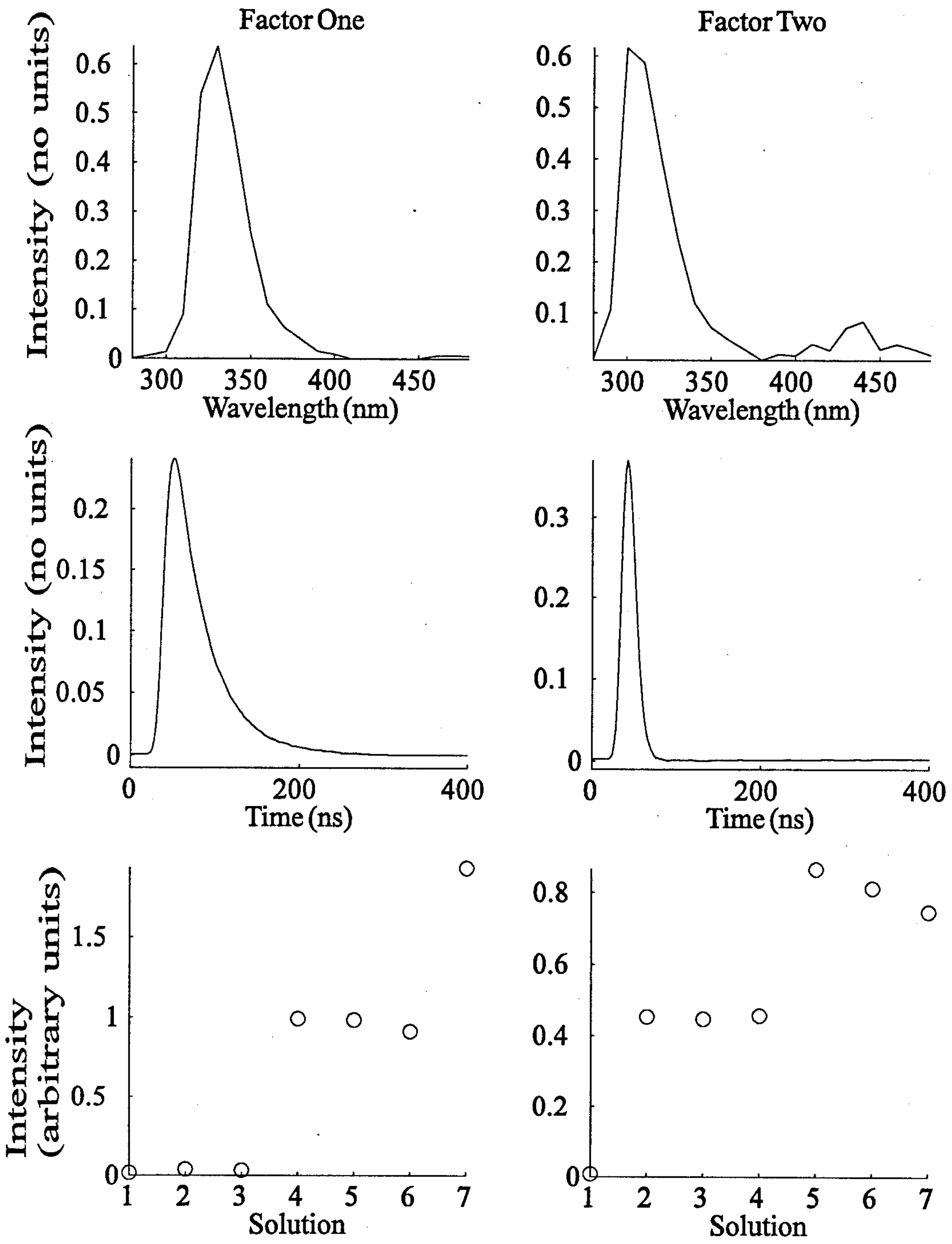


Figure 5.4. Rank 4 3M-NNALS determined factors for Data Set One.

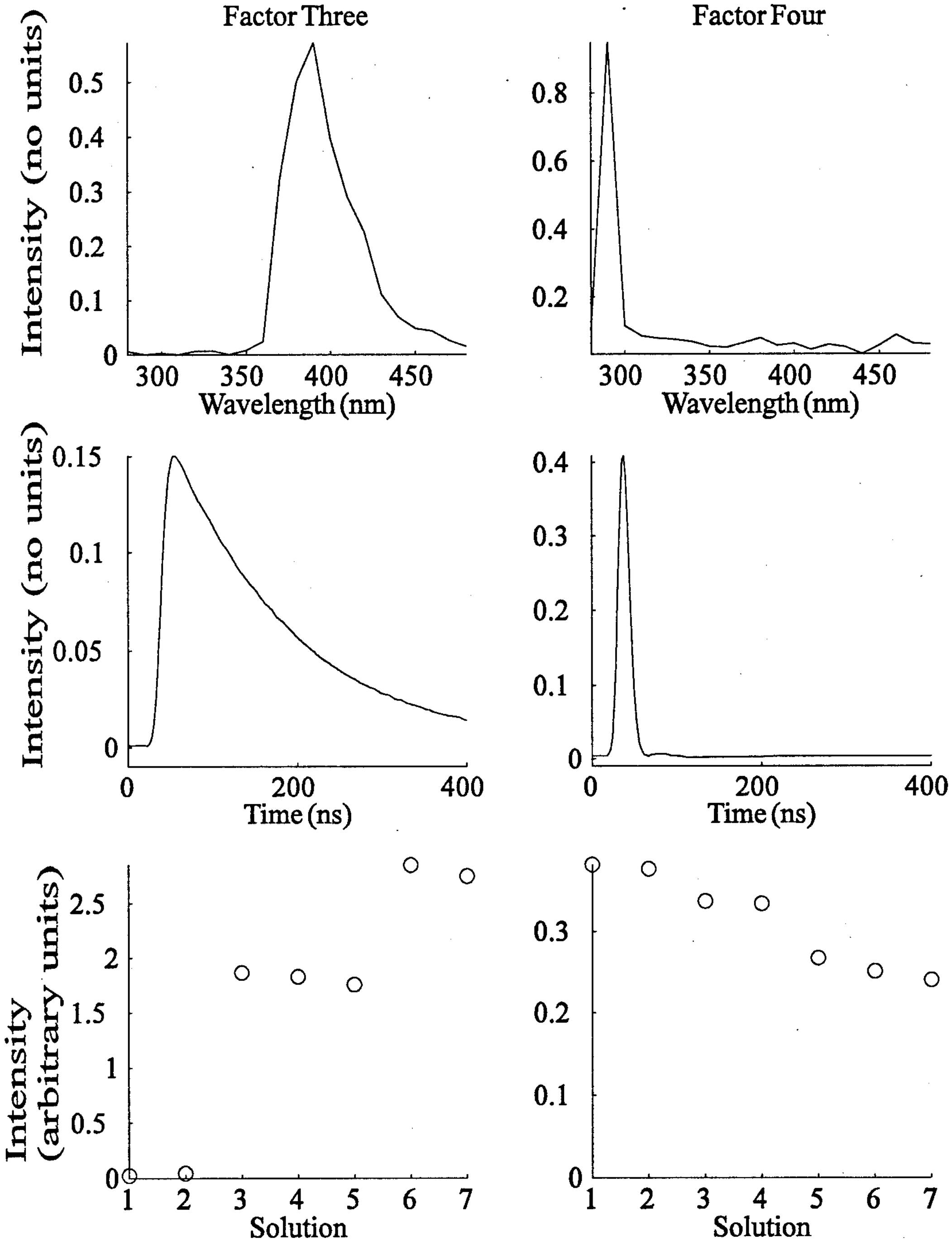


Figure 5.4. Continued.

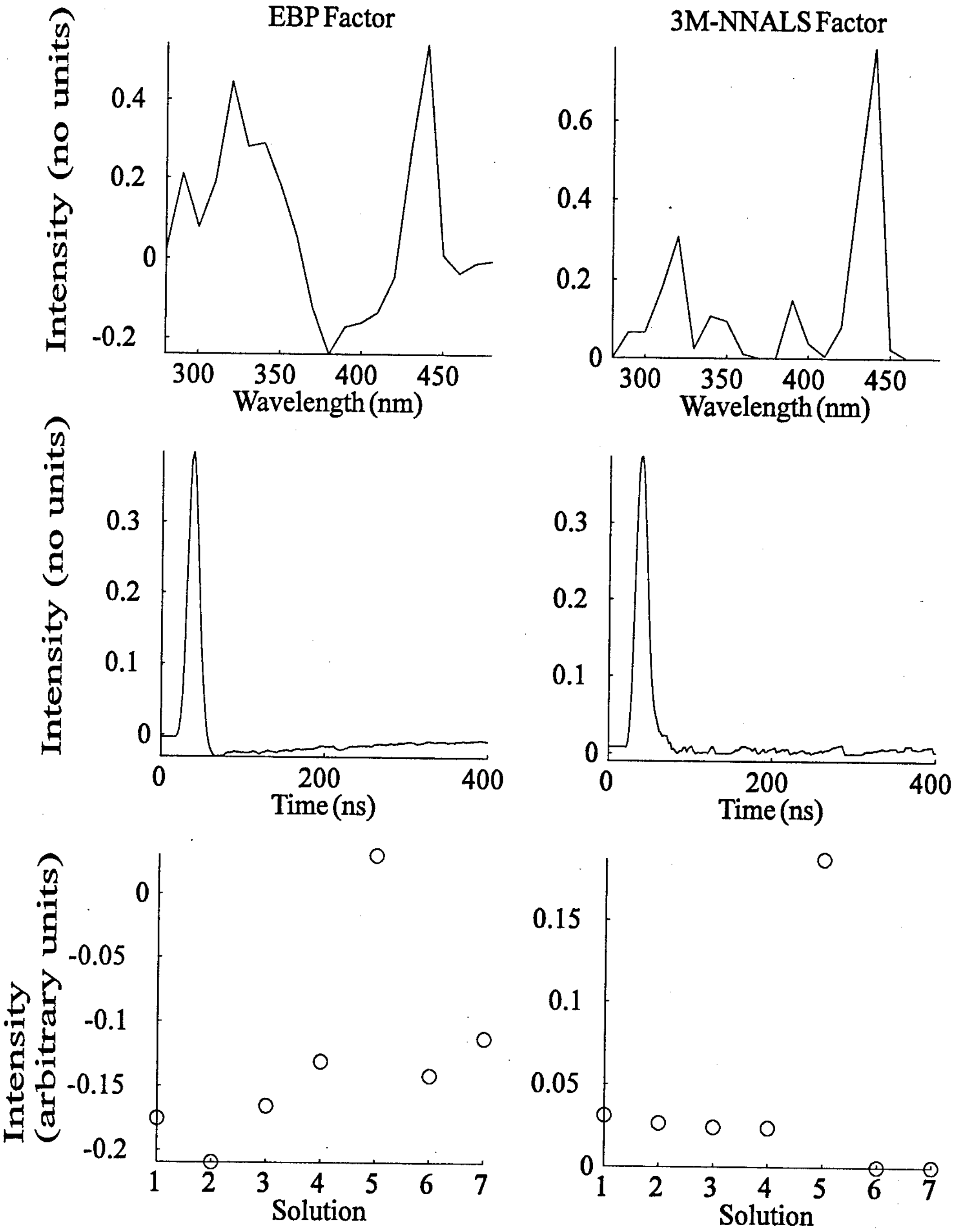


Figure 5.5. Rank 5 decomposition of Data Set One - Fifth factor.

The 3M-ALS sequence required 854 iterations to converge for the rank 5 resolution when starting with the factors from the EBP, whereas the 3M-NNALS required only 116 iterations to converge, starting from the same factors.

The other resolutions, ranks three and six, all produced factors recognizable as the three analytes after 3M-ALS or 3M-NNALS. The rank three resolution coalesced the fluorene and Raman factors, which are the most similar of the four chemical components in wavelength and in fluorescence decay time. In the TLD for assumed rank six, all of the factors from the rank five TLD were obtained. In addition, a small magnitude, structureless, short lifetime noise factor arose, which is the expected result when a decomposition of rank higher than the actual array rank is sought.

5.1.1.1.3 Comments on TLD of Data Set One

When confronted with relatively simple data, such as that contained in Data Set One, the EBP does a very good job of extracting the profiles of the major components in all three modes. 3M-ALS is still needed to make further refinements in the profiles, evidenced by the improvement in the fluorene factor following 3M-ALS. The EBP's ability to extract weaker components is not as good as 3M-ALS. One may even have overlooked the mysterious fifth component if 3M-ALS had not been performed.

3M-NNALS, on the other hand, did not represent a significant improvement over the 3M-ALS result with respect to the quality of the factors. Nonetheless, 3M-NNALS is beneficial because it provides a faster route, sometimes substantially faster, to convergence. It is not clear why the rank five 3M-ALS solution took so long to converge, but the 3M-NNALS solution apparently avoided that problem.

3M-NNALS is undoubtedly more efficient than 3M-ALS at avoiding physically meaningless regions of hyperspace. 3M-ALS allows the solution to wander through hyperspace in search of a minimum. We have observed that quite often two corresponding profiles for a factor (time and wavelength, say) will be inverted. It is possible that the phenomenon of "swamps" described by Mitchell and Burdick[63] can be avoided by 3M-NNALS; surely, it precludes two factor degeneracies.

Rank estimation is another area requiring comment. Considering the Raman scatter and the contaminants as independent factors, then Data Set One constitutes a rank five 3-array. The best estimates given for the rank seem to be from the factor indicator function and Malinowski's *F*-test performed on the summed WTMs. Other than the estimate for Solution 1, there were no rank five estimates for the WTMs. Even Solution 5, where the contaminant was most evident, the rank estimate was "low." Summation provides some smoothing or signal averaging of the data, which may be enough to allow recognition of weaker factors, particularly those associated with processes such as solvent Raman scattering, which is represented in all the WTMs of the data array.

5.1.1.2 Data Set Two: A TREEM

Data Set Two was measured on a single aqueous solution that contained fluorene, phenanthrene, naphthalene, and carbazole. The solution was prepared by adding small amounts of stock solutions of the individual components in methanol to 3 mL of water in a quartz cuvette until a suitable signal level was reached. A TREEM consisting of four WTMs at excitation wavelengths of 287 nm, 290 nm, 295 nm, and 300 nm was measured. The data interval and range were 2 ns over 250 ns along the time mode and 2.5 nm from

305 nm to 405 nm along the wavelength mode. Other experimental conditions were identical to those used for the acquisition of Data Set One. The WTMs that comprise the TREEM for Data Set Two are presented in Figure 5.6.

Figure 5.6 reveals that there is much less variation from one WTM to another in comparison to Data Set One. Note that the variation in these WTMs arises solely from differences in the spectral response due to changes in the excitation wavelength. The concentrations of the fluorescing species do not change from one WTM to the next. One might expect the estimation of the rank of this 3-array to be more difficult than that of Data Set One, given the similarity of the WTMs.

5.1.1.2.1 Rank Estimation of Data Set Two

Rank estimation for Data Set Two was performed in an identical manner as Data Set One. The results of these analyses are presented in Table 5.3.

Based on *a priori* knowledge of the solution, the best estimates of rank for this data are given by the wavelength-mode autocorrelation and Malinowski's *F*-test. The factor indicator function was not a reliable estimate of rank estimation for any of the WTMs in this data set.

Table 5.3. Rank estimates for Data Set Two.

Excitation Wavelength (nm)	Rank[1]	IND[2]	*F*-test[3] (5 %)	C(**u**(λ))[4]	C(**u**(t))[4]
287	4	6	4	3	10
290	4	9	4	3	8
295	4	8	4	3	9
300	4	5	4	4	11
Σ(WTMs)	4	10	5	4	10

[1] Rank is equal to the number of emitting species.
[2] Malinowski's factor indicator.
[3] Malinowski's *F*-test for significance above the 5% level.
[4] Autocorrelation coefficient, values above +0.5.

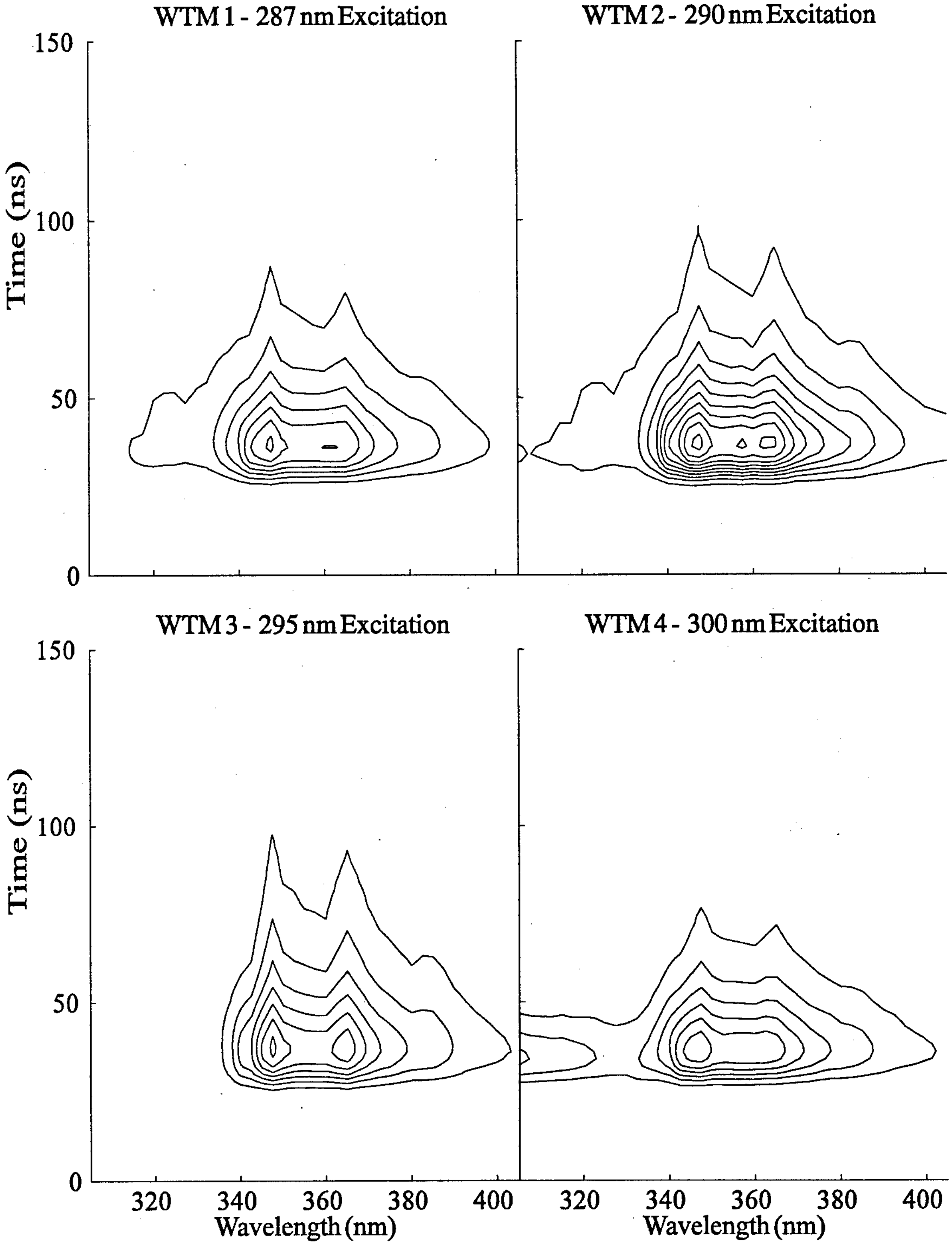


Figure 5.6. TREEM of Data Set Two. Contours represent levels of equal fluorescence intensity. The weakest intensity is represented by the outermost contour in each WTM; it has a value of 0.1 (arbitrary units). Intensity increases by 0.1 for each successive contour.

Raman scatter was not considered as a contributor to these data. Because the excitation wavelength was varied, the water Raman scattering would appear at a different position in each WTM. If the Raman scatter were included, the actual rank would be higher by a value of four. However, since the concentrations of the fluorophores was elevated to optimize the signal, the Raman scatter was diminished substantially by competition for incoming photons. The intensity of the Raman scatter decreases as the concentrations of the luminescent species are increased. This behavior can be observed in Figure 5.4, Factor Four, Solution-mode.

5.1.1.2.2 TLD of Data Set Two

EBP decomposition was performed on Data Set Two in the same manner as for Data Set One. Decompositions were performed for assumed ranks two through six. Complex (imaginary) factors resulted in one set of the rank three factors and in both sets of the rank four, five, and six factors (Note: this EBP produces two sets of factors, one corresponding to eigenanalysis on the G matrices, the other corresponding to eigenanalysis on the transpose of the G matrices). Complex factors can arise during eigenanalysis on nonsymmetric square matrices (see Appendix C), which is generally the case in the approaches required here (see equations C.12 and C.13). Li *et al.*[72] devised an algorithm to eliminate complex eigenvalues during EBPs based on similarity transforms.

The results of the rank four EBP utilizing the similarity transform method of Li *et al.*[72] are displayed in Figure 5.7. None of these factors can clearly be assigned to any one species. They appear to be linear combinations of the individual component spectral and time profiles, that is, they are rotated together. Each of the various rank resolutions

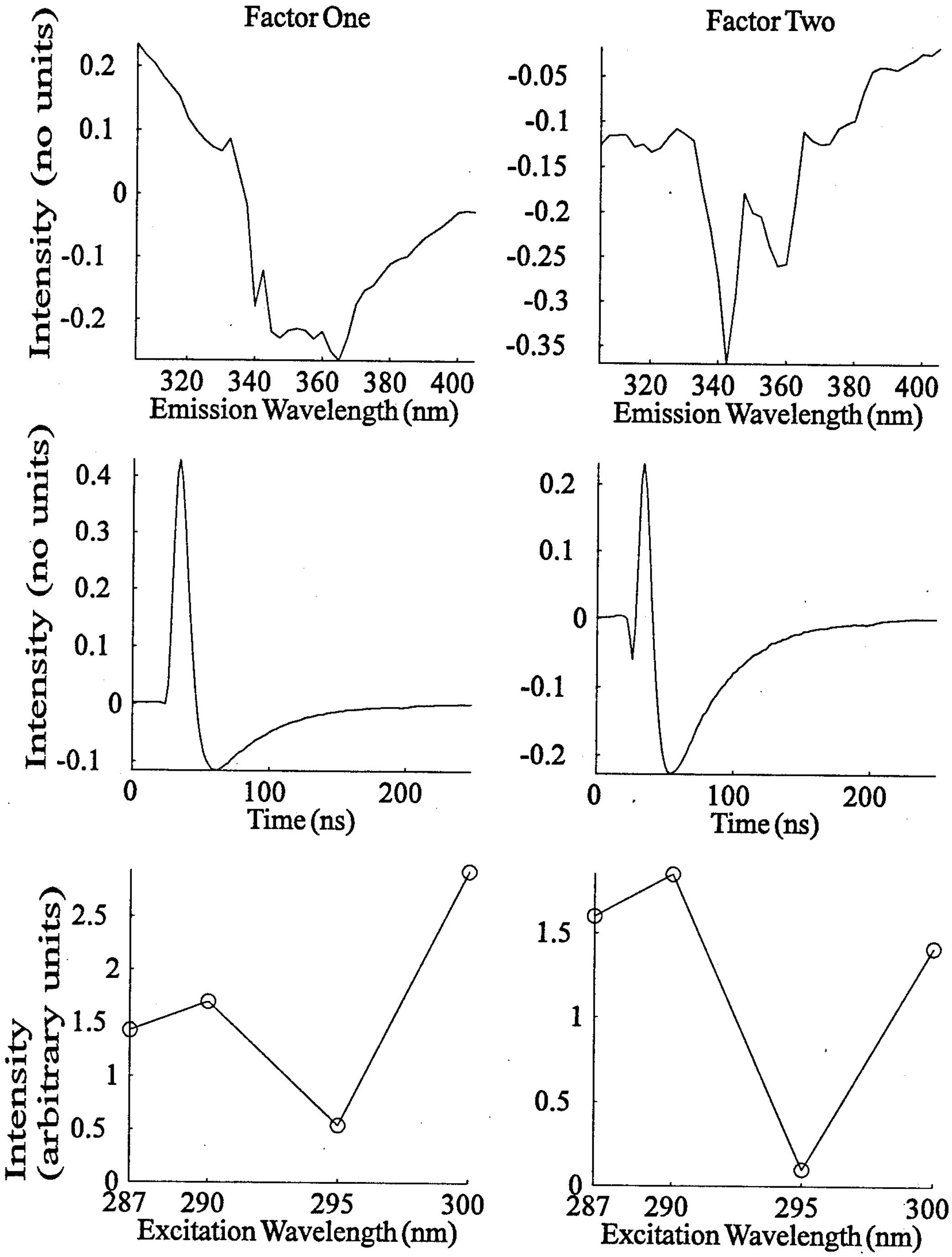


Figure 5.7. EBP determined factors for Data Set Two.

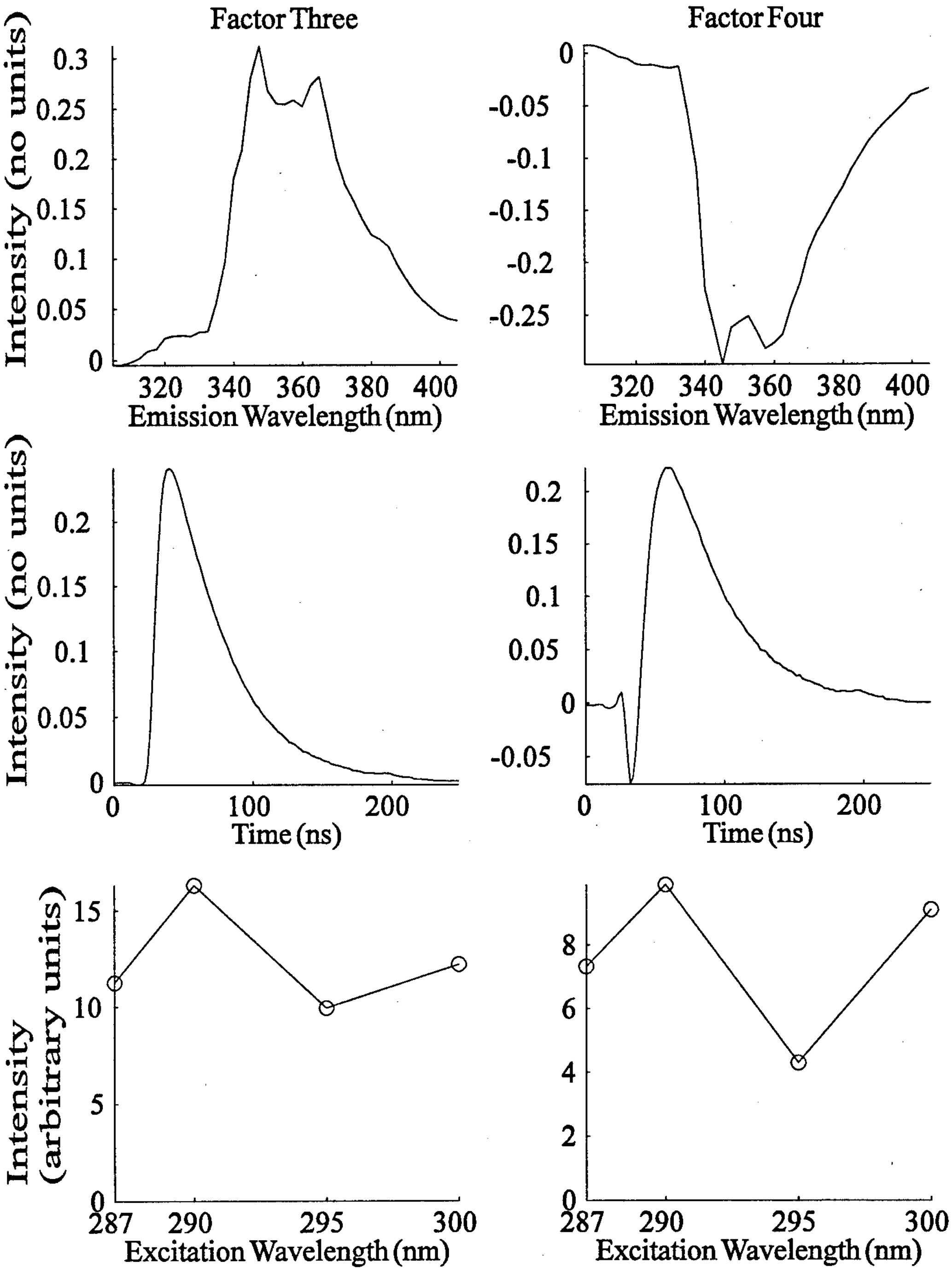


Figure 5.7. Continued.

resulted in factors similar to the rank four resolution. 3M-ALS is obviously required in this case.

The rank four 3M-ALS decomposition of Data Set Two is exhibited in Figure 5.8. The respective factors from Figure 5.7 were used as starting vectors for 3M-ALS. Convergence occurred after 3846 iterations. The improvement in the quality and appearance of the factors is noteworthy.

Based on the emission wavelength-mode and time-mode factors, Factors One through Four in Figure 5.8 can assigned as fluorene, carbazole, naphthalene, and phenanthrene, respectively. Fluorescence lifetimes are 6 ns, 11 ns, 30 ns, and 32 ns for Factors One through Four, respectively. They compare favorably with 7 ns for fluorene in water,[30] 15 ns for carbazole in ethanol,[73] 36 ns for naphthalene in water,[30] and 33 ns for phenanthrene in water.[30] Lifetimes were calculated with the phase plane method with a separate measurement of scattered laser excitation to provide $E(t)$.

The emission wavelength-modes for Factor Three and Factor Four indicate that the two are rotated together. Note that the short wavelength portion of Factor Four has a very similar structure to that in Factor Three. The similarity of the lifetimes and the effects of experimental noise are each likely contributors to the vector rotation. Even if the 3M-ALS process is allowed to continue until stricter termination criteria are met, the result is essentially the same. The need for constraints during the 3M-ALS sequence is more evident for Data Set Two than for Data Set One.

3M-NNALS was performed on the data using the factors in Figure 5.7 as starting points. The sequence converged quickly, taking only 337 iterations. Figure 5.9 contains

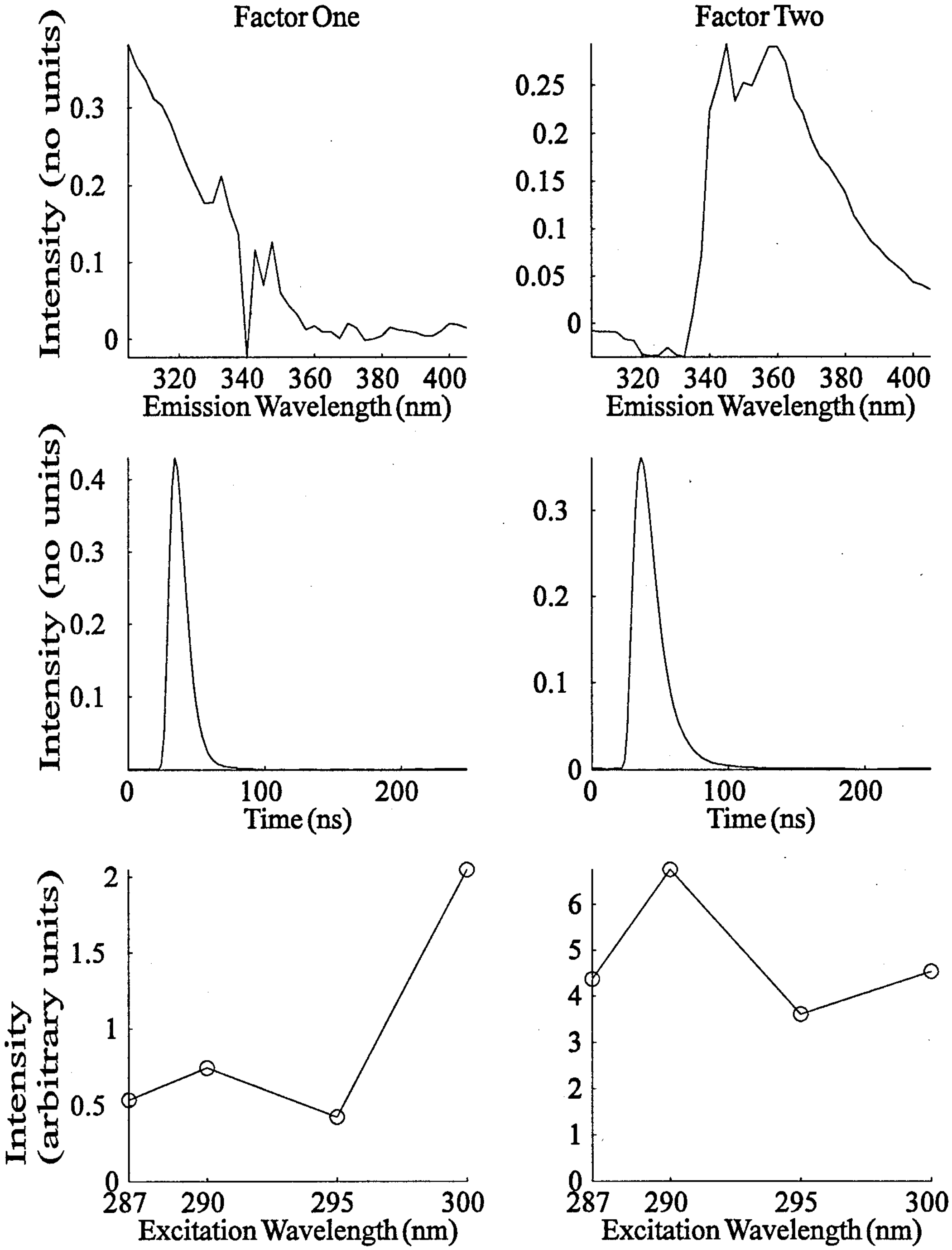


Figure 5.8. 3M-ALS determined factors for Data Set Two.

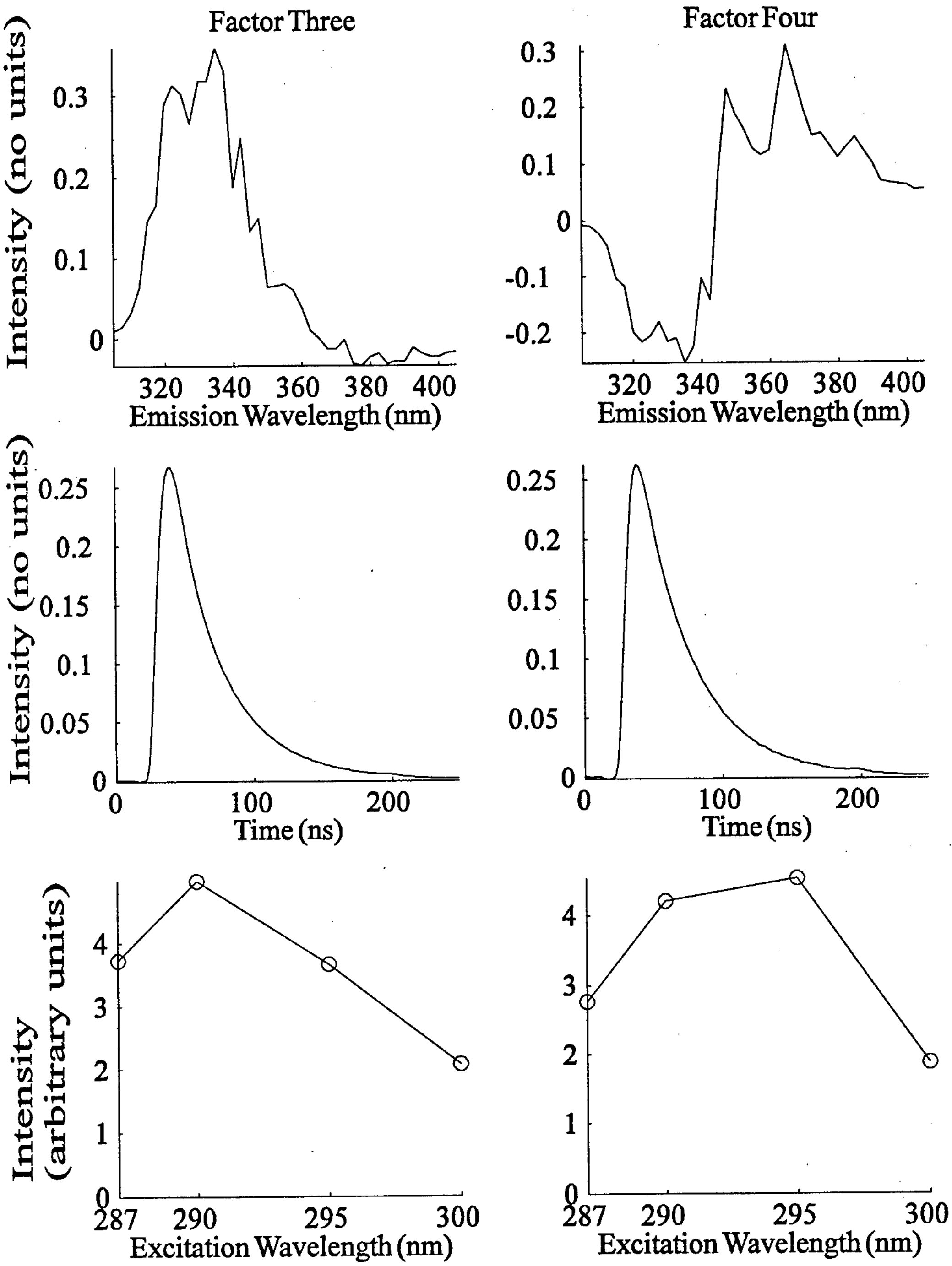


Figure 5.8. Continued.

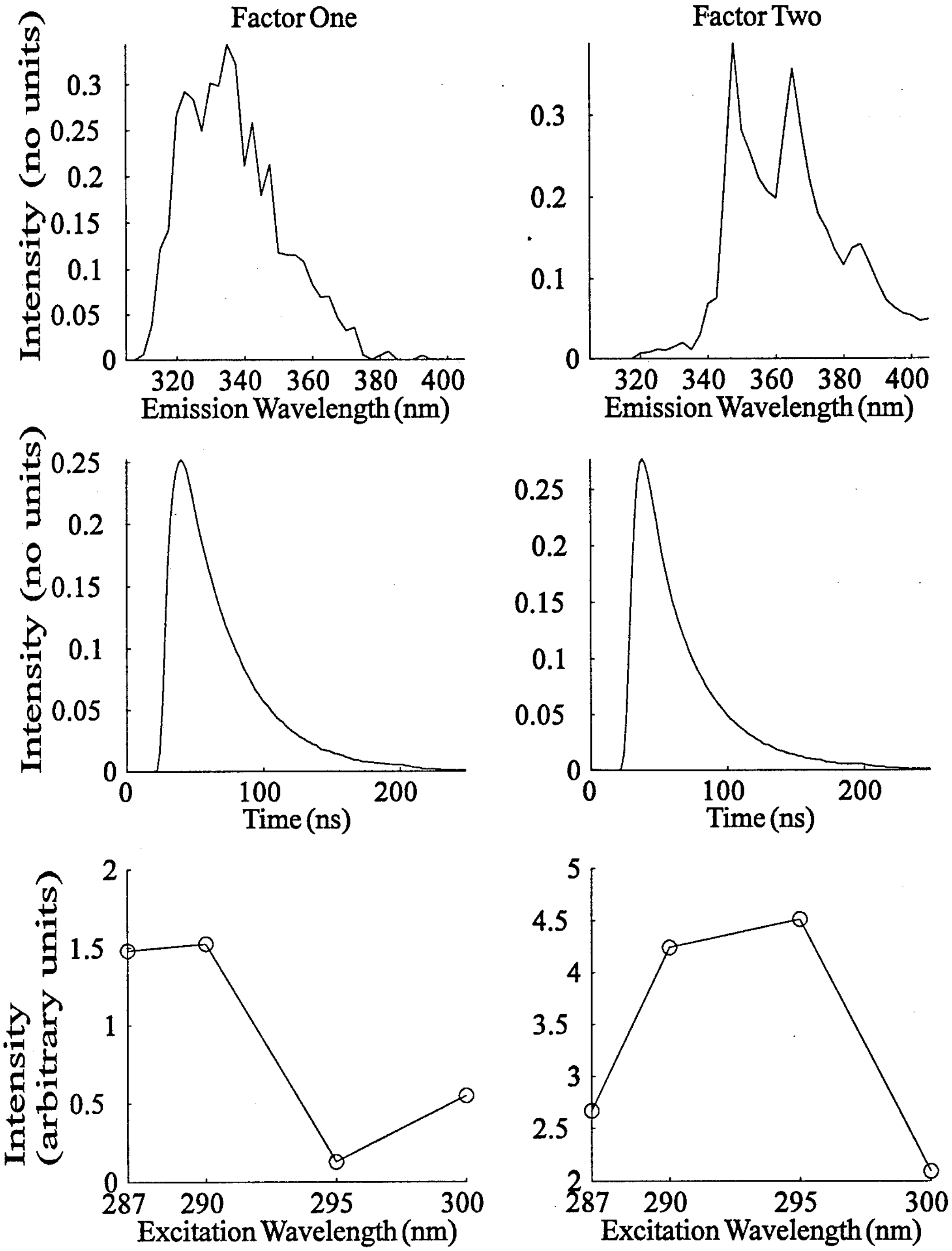


Figure 5.9. 3M-NNALS determined factors for Data Set Two.

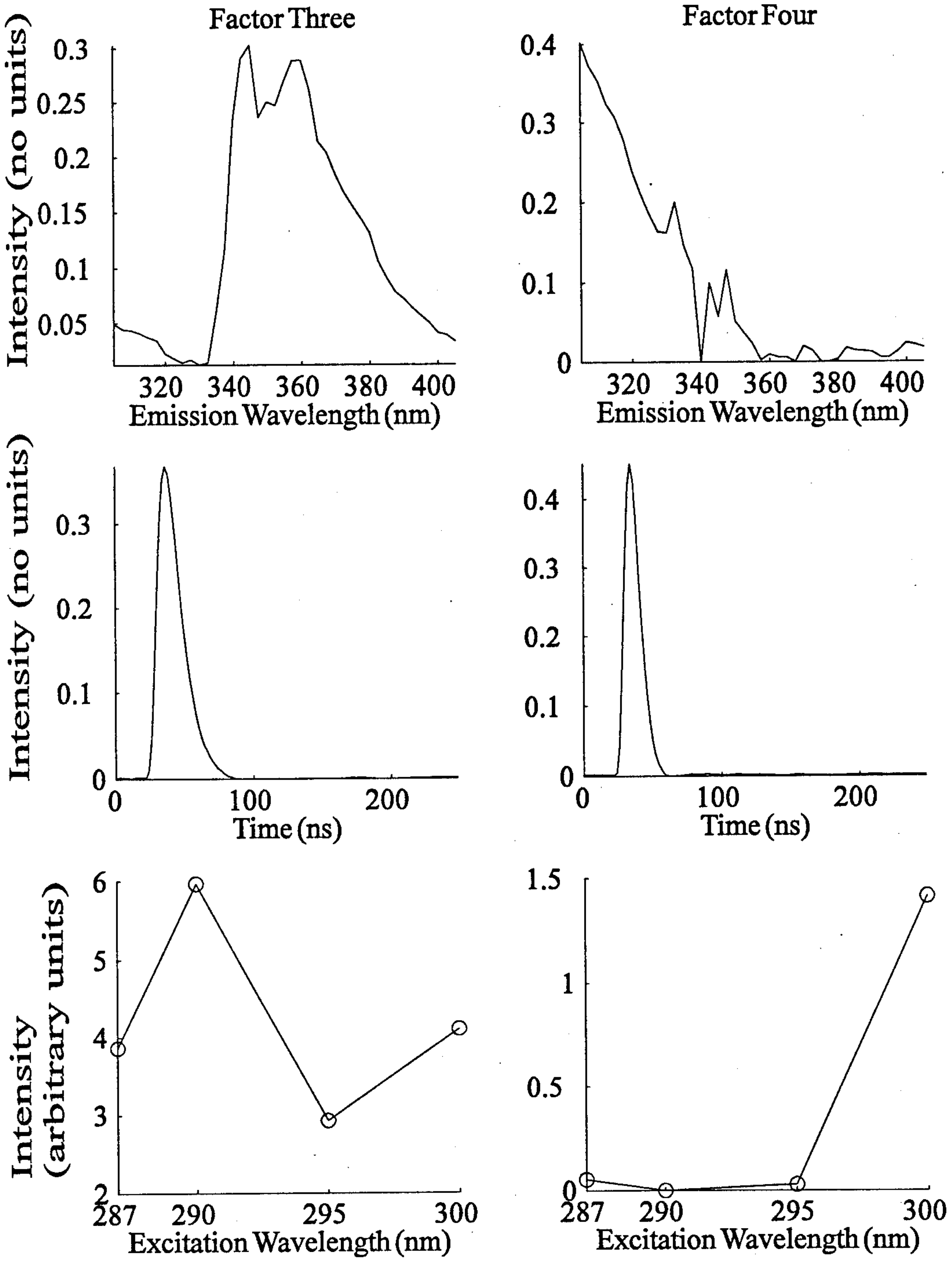


Figure 5.9. Continued.

the rather impressive results. Our assignments for Factors One through Four are naphthalene, phenanthrene, carbazole, and fluorene, respectively; the computed lifetimes are 34 ns, 29 ns, 10 ns, and 5 ns, respectively. Note that the factors have not been renumbered from those given by the algorithms to better match them up with real factors.

The 3M-NNALS factors are more realistic than the EBP factors and are also an improvement over the 3M-ALS result since the rotation of the naphthalene and phenanthrene factors is eliminated. The emission mode 3M-NNALS factors are compared with fluorescence spectra measured on a spectrofluorimeter (Spec 2T2 Fluorolog) in Figure 5.10. The agreement is excellent, particularly since the data are not corrected for detector response in either case.

The excitation wavelength modes have received little attention thus far. This is because there were so few measurements in this mode; also the WTMs were not corrected for laser power variation as a function of excitation wavelength. This mode provided variations in fluorescence intensity for the luminescent species, although we acknowledge the potential of excitation wavelength as an additional identifying agent.

Resolutions for the other assumed ranks were less satisfactory than the rank four results. The rank three resolution produced factors similar to Factors One, Three, and Four in Figure 5.8. The residual hypersurface of the rank three 3M-NNALS decomposition apparently contained MLO. Some starting vectors led to two reasonable factors while others led to only one. This is not too surprising since the rank three resolution is rank deficient.

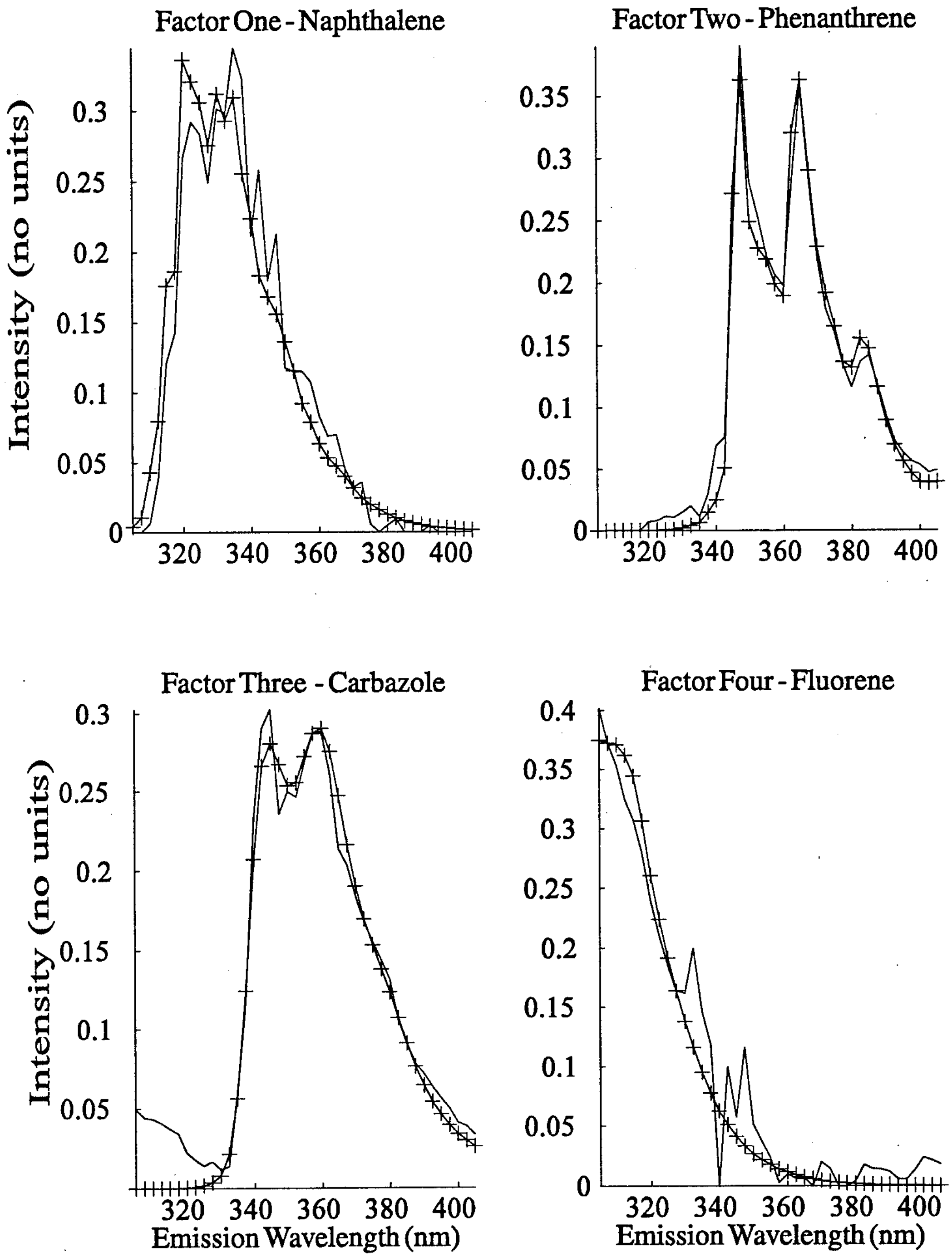


Figure 5.10. Comparison of resolved factors to pure species spectra. Pure component spectra are the crossed solid lines. Emission wavelength-mode factors are solid lines.

The 3M-NNALS rank five decomposition produced reasonable factors for fluorene, phenanthrene, and carbazole. The naphthalene factor appeared to be rotated with the phenanthrene factor. However, the decomposition did not generate the expected noise factor. Instead, a fifth factor emerged with a lifetime and spectrum similar to carbazole.

The rank six decompositions failed to converge after more than eight hours of computational time for both 3M-ALS and 3M-NNALS.

5.1.1.2.3 Comments on Data Set Two

Data Set Two represents a real challenge for three-mode decomposition, especially for EBP and 3M-ALS. The rank four decomposition by 3M-NNALS is highly encouraging. Meaningful profiles were recovered in the spectral and time domains, and the number of iterations required to reach convergence was reduced over the unconstrained procedures.

In all probability, the most significant impediment to a good 3M-ALS solution is the presence of real factors with very similar fluorescence lifetimes. The prevention of factor rotations, such as those which occurred between the naphthalene and phenanthrene factors, underlies the potential of 3M-NNALS and similar analyses, such as restricted Tucker models.[37]

The complex solutions to the EBP analyses point to what may be an instrumental challenge in the acquisition of this type of time-domain data. Sanchez and Kowalski[36] suggested that complex solutions in the EBP results are a consequence of deviations from trilinearity. We have observed that the dye laser pulse duration can vary with wavelength. If it does, then the fluorescence decay profiles of components whose lifetimes are not much longer than the laser pulse duration are similarly affected, and the WTMs are distorted. One way to overcome this problem in the future would be to deconvolve the data before three-mode decomposition, for example, by the exponential series method for fitting multiexponential decays.[74] Alternatively, one could attempt modifications to the dye laser itself to make the pulse durations as constant as possible.

Predictions of the ultimate potential for rank estimation from TREEMs is difficult with the limited data presented here. *A priori* information and the quality of the rank four 3M-NNALS decomposition suggest that the data are truly rank four. However, the emergence of an apparent fifth factor in the rank five decomposition and the abundance of complex EBP solutions hint at a rank ambiguity. The estimate of rank five from the F-test may be further support for such a conclusion. Even so, the F-test operating on the summed WTMs still provides a reasonable place to start.

It is unclear why the rank six decomposition failed to converge. This behavior is unusual, especially for the 3M-NNALS, and we cannot offer an explanation at this time.

5.1.2 Global Analysis of Three-mode Data

The power of global analysis lies in its ability to simultaneously model some physical observable across a number of independent measured data sets. Thus, it provides excellent opportunities for signal averaging and can describe a large set of data with relatively few parameters.

5.1.2.1 Global Analysis of Data Set One

Data Set One was introduced in Section 5.1.1.1. There is a particular aspect of this data set which allows it to be easily analyzed using global analysis. The Raman scatter in the first WTM provides an excellent system response function, $E(t)$, for convolution. Since global analysis essentially performs curve fitting using the convolution of $E(t)$ and the exponential decay of the fluorophore, $y(t)$, collecting a high quality $E(t)$ is crucial.

Global analysis was conducted using the Levenberg-Marquardt search algorithm which is packaged in the *Optimization Toolbox* for Matlab®. Convolutions were performed by numerical integration of equation 2.11 using a polynomial integration program written in this laboratory. After each set of lifetime parameters was estimated, a least squares fit of the data was performed to extract the solution-mode and wavelength-mode parameters. The 3-array was unfolded as an ${}_i\mathbf{X}_{(jk)}$ matrix (see Appendix B) with the *i*-mode representing time, the *j*-mode representing solution, and the *k*-mode representing emission wavelength. Termination was based on reaching a minimum in the norm of the residual matrix.

To model these data based on five components, global analysis was conducted using a zero lifetime and four nonzero lifetimes. The excitation profile was used without convolution as a zero lifetime component to accommodate scatter. Unfortunately, a global analysis solution to a five-component problem could not be reached. Attempts generally resulted in two very similar lifetimes, usually at the extremes. Consequently, solution of the least squares problem leading to the solution-wavelength-mode was poorly defined (i.e., if $\mathbf{A}$ represents the time-mode matrix, $(\mathbf{A}'\mathbf{A})^{-1}$ was nearly singular), and it would have the properties similar to a two-factor degeneracy in 3M-ALS. A four-factor global analysis solution was pursued, in light of the failure of the five-factor model.

The results of a four-factor (a zero lifetime and three nonzero lifetimes) global analysis are exhibited in Figures 5.11 through 5.14. Components represented by the factors are in

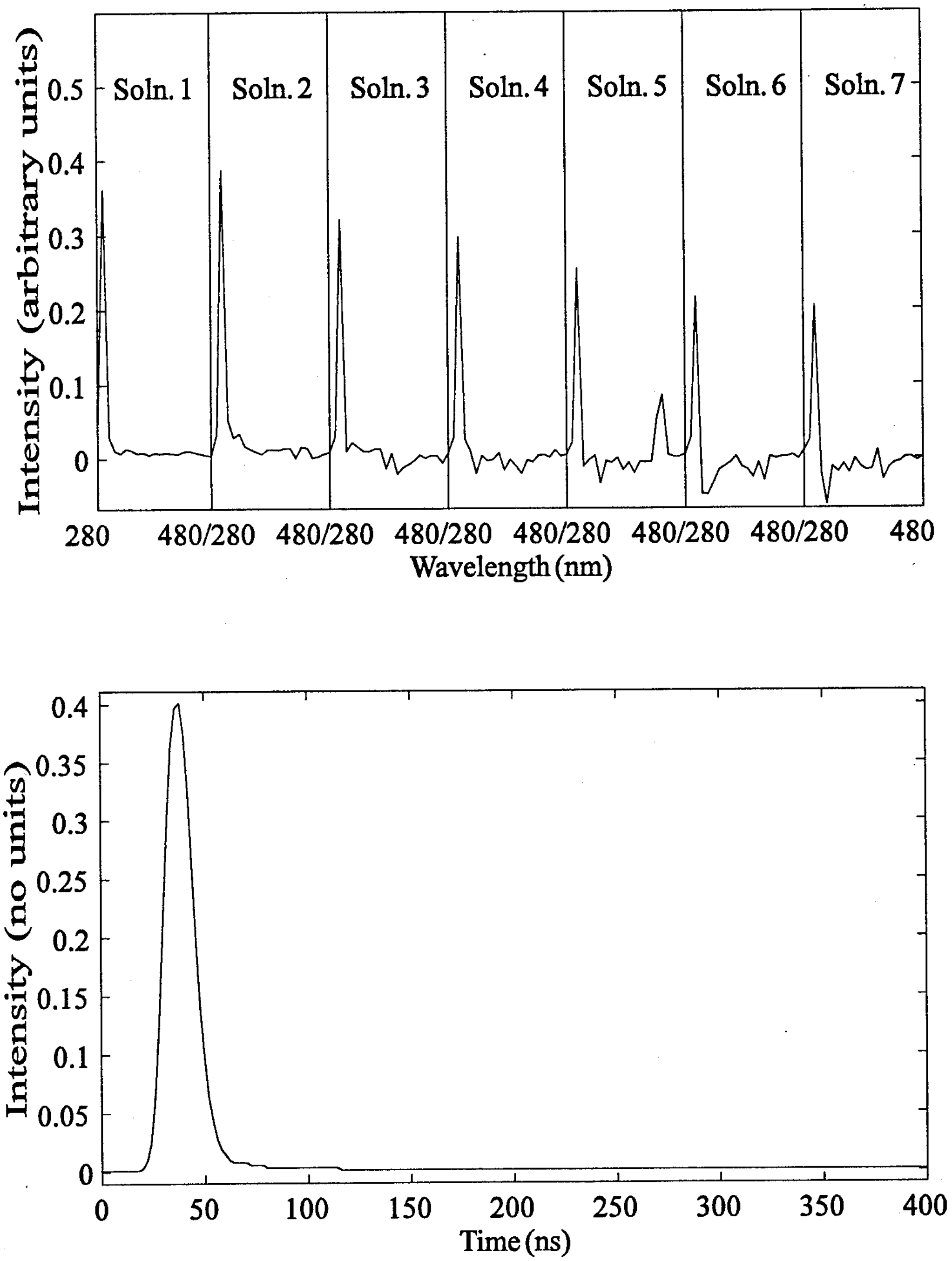


Figure 5.11. Factor One global analysis result for Data Set One. The upper plot is the least squares estimate of the solution-wavelength-mode factors. The lower plot is the instrument response function.

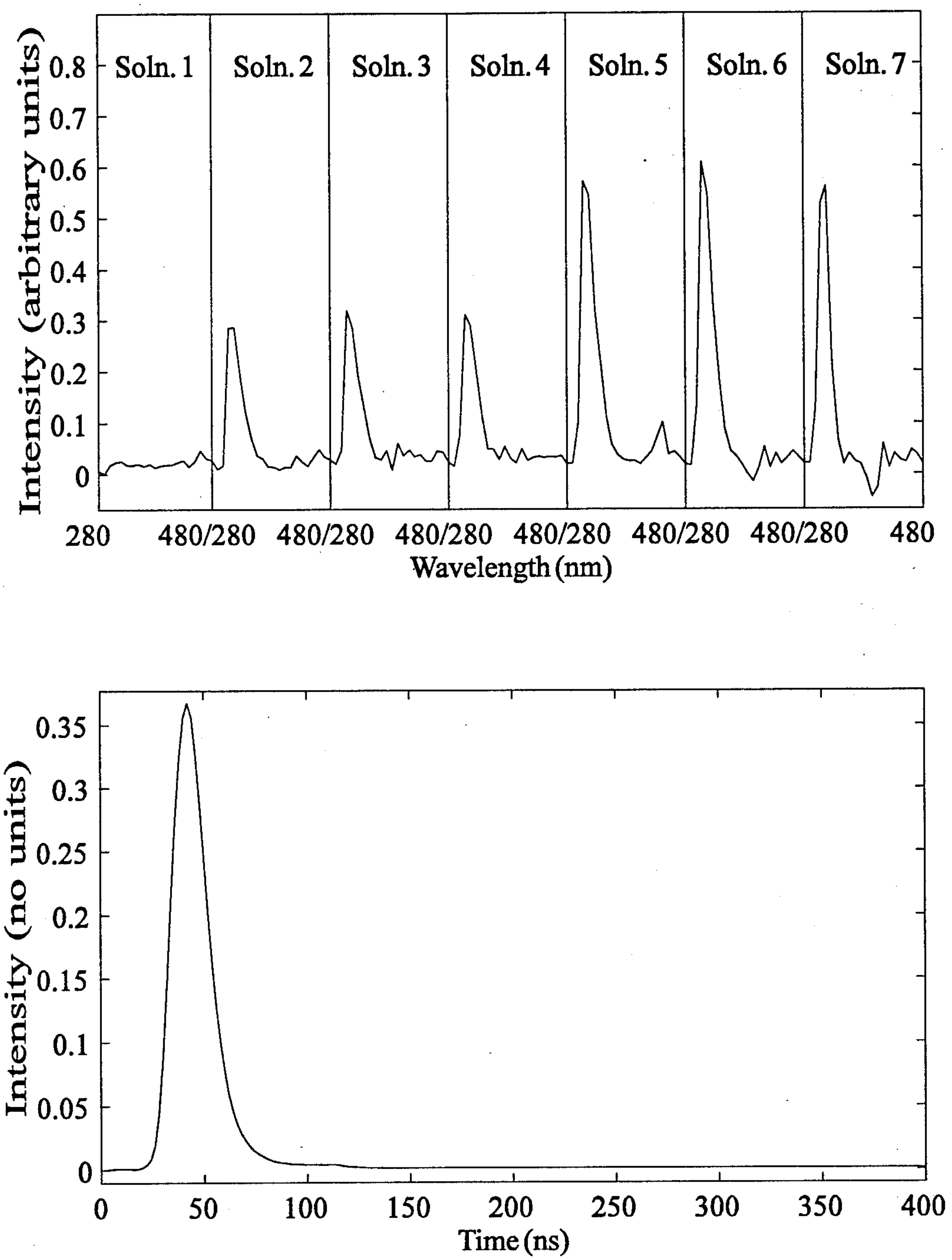
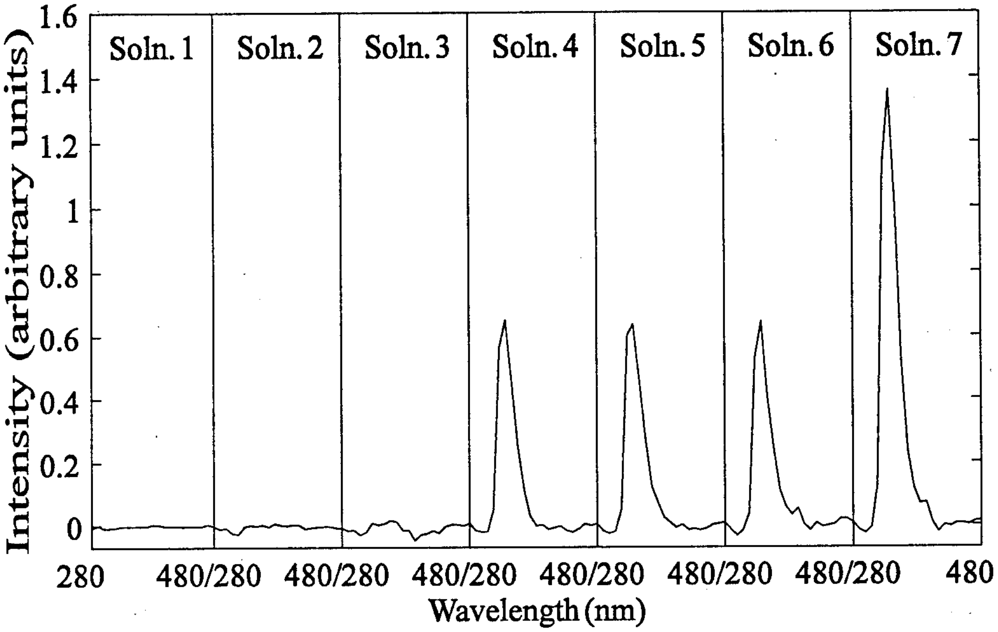


Figure 5.12. Factor Two global analysis result for Data Set One. The upper plot is the least squares estimate of the solution-wavelength-mode factors. The lower plot is the convolution of the instrument response with the decay for a 6.2 ns lifetime.



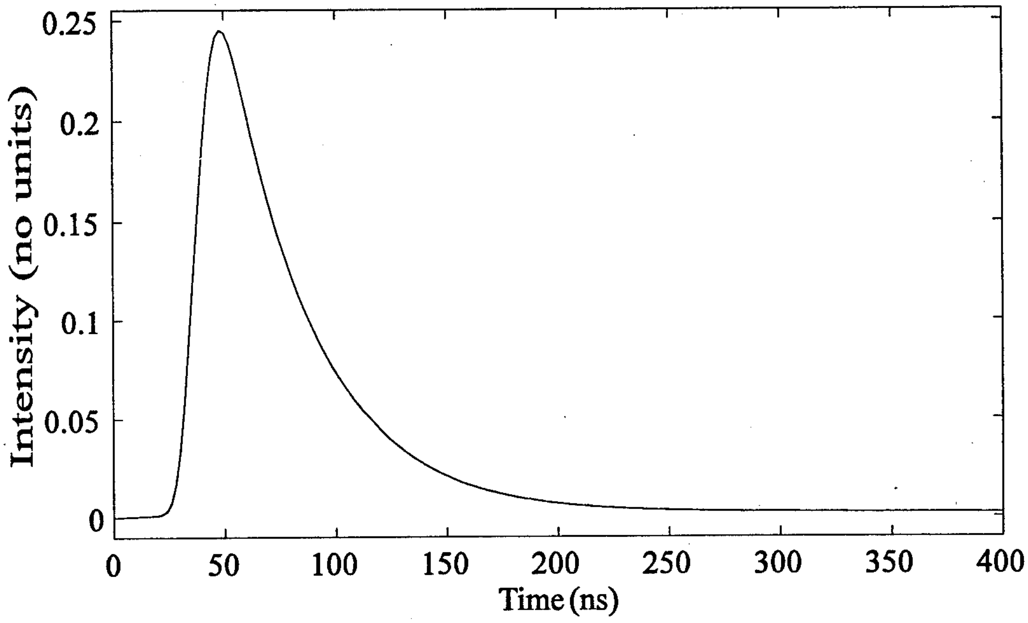


Figure 5.13. Factor Three global analysis result for Data Set One. The upper plot is the least squares estimate of the solution-wavelength-mode factors. The lower plot is the convolution of the instrument response with the decay for a 36.4 ns lifetime.

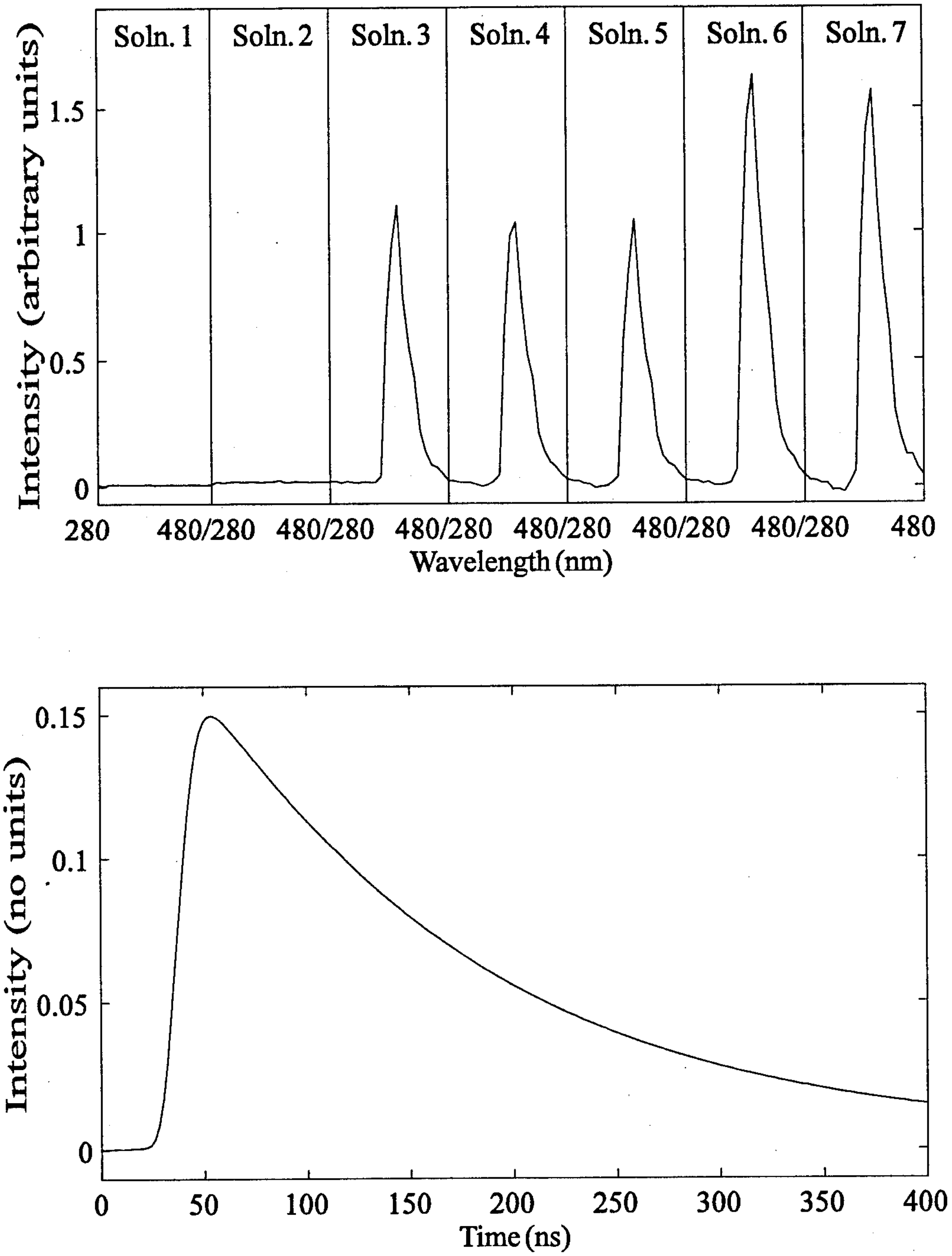


Figure 5.14. Factor Four global analysis result for Data Set One. The upper plot is the least squares estimate of the solution-wavelength-mode factors. The lower plot is the convolution of the instrument response with the decay for a 133 ns lifetime

order: Raman scatter, fluorene, naphthalene, and pyrene. The lower portion of the figures contains the time-mode part, and the upper portion contains the least squares estimate of the solution-wavelength-mode. In Figure 5.11, in the region of Solution Five, one can see the contribution of the contaminant. It can be seen to a lesser degree in the same region in Figure 5.12. Taken together, the two signals can be fit to a lifetime of 2.8 ns, almost identical to the 3M-ALS result. Why this factor could not be extracted is not clear. Perhaps 3M-ALS and 3M-NNALS are more robust techniques for extracting factors that make relatively small contributions to the data.

The results for the remaining factors are quite good. The lifetimes for the fluorene, naphthalene, and pyrene, 7 ns, 36 ns, and 133 ns, respectively, are in good agreement with other measurements performed in this laboratory (see Section 5.1.1.1.2). They differ from those obtained from 3M-ALS because global analysis enforces the model in equation 2.11 during the fitting process, while 3M-ALS fits the data to the trilinear model without regard to lifetime.

Another noteworthy point about these fits is the behavior of the scatter or zero-lifetime factor. As one moves from Solution One to Solution Seven in the top portion of Figure 5.11, there is an increasing *negative* contribution of the zero-lifetime factor. This effect becomes markedly more prominent when a discrete convolution is performed in place of a numerical integration. This effect has been observed in this laboratory on several occasions.

Commonly, the problem is treated as a time shift phenomenon and is easily accommodated in this way. However, it has also been observed in this laboratory that a small time shift has the same effect as a short lifetime contribution. One possible cause of this behavior is the numerical treatment of the data; another is that the response of PMT may be nonlinear on the rising edge.

Carefully comparing the time-mode profiles in Figure 5.4 with those in Figures 5.11, 5.13, and 5.14, a small difference in the rising edges can be detected. The global analysis profiles consistently rise faster than the 3M-NNALS profiles. This may be a contributing factor to the problem of the negative signal contribution; and it may have the same cause. All of these matters are under investigation.

5.1.2.2 Global Analysis of Data Set Two

Efforts to perform global analysis on Data Set Two using the techniques described in Section 5.1.2.1 were unsuccessful. Three was the maximum number of lifetimes which could be determined. Attempts to obtain four lifetimes consistently failed due to difficulties identical to those encountered in the five-lifetime analysis of Data Set One.

The three lifetime global analysis did not produce clearly recognizable factors. It also was extremely sensitive to starting point. This result was not too surprising, given the results of the rank three decomposition from 3M-ALS and 3M-NNALS.

Considering the success of the rank four 3M-NNALS solution for this data, we wrote and performed a three-mode global analysis algorithm. The algorithm searched lifetime parameter space in a nonlinear fashion, however, for each new lifetime, the wavelength-mode and solution-mode were solved by a two-mode NNALS routine, rather than solving them together.

The four-lifetime three-mode global analysis produced excellent results. Both spectral and temporal profiles were nearly indistinguishable from the results in Figure 5.9. Lifetimes extracted were 5 ns, 11 ns, 35 ns, and 38 ns, for the fluorene, carbazole, phenanthrene, and naphthalene factors, respectively.

This new method of global analysis has produced encouraging results. Its major drawback is that it requires extensive iteration.

5.2 Profile Extraction of Fuel Fluorescence

An evaluation of the performance of a prototype of the ROST™ system in comparison to a nitrogen laser-based system designed for environmental analysis was performed as part of the DoD Site Characterization and Analysis Penetrometer System (SCAPS) program. The testing was conducted at the Naval Command, Control Ocean Surveillance Center, Research, Development, Test and Evaluation Division (NRaD) during February, 1994. During the evaluation, numerous data sets were acquired to assess the capability of the two fluorescence systems to detect and identify fuels on soil matrices and to quantitate contamination levels. The following four fuels were selected for the test on the basis of their widespread distribution at military sites throughout the country: diesel fuel-marine (DFM), summer grade diesel fuel (DF), unleaded gasoline (UG), and JP-4 jet aircraft fuel (JP4).

In one part of the evaluation, participants were challenged to test the ability of the systems to identify unknown fuels. Hence, four fuels were placed onto three different soil matrices in different concentrations. Soil matrices also were chosen for their diversity; the first was sand, nonabsorbent with a light background; the second was from Columbus AFB, MS, (CAFB) and was highly absorbent and nonreflective; the third was from China Lake NAS, CA, (CLNAS) and was also absorbent and nonreflective. Individual fuel-soil mixtures were loaded into sets of soil sample holders fitted with a sapphire window. The loaded sample holders rotated during measurement to minimize photolysis while spectra were simultaneously measured with the ROST™ prototype and the nitrogen laser systems. Spectral responses of the 12 combinations of the four fuels on the three soils at a concentration of 3000 ppm served as a calibration set. As a blind test of the capabilities of the two systems, 24 unknowns were prepared in concentrations of 1000 or 10,000 ppm and measured.

The fluorescence data collected by the ROST™ prototype were acquired for 290 nm excitation. Twenty-one emission wavelengths evenly spaced between 300 nm and 500 nm were monitored. Decays were measured at 1 ns intervals over a 120 ns range. Fiber-optic light delivery and collection occurred over a 50 m long silica-clad silica two-fiber probe. The core diameter was 365 µm. A representative WTM for each of the fuel products is displayed in Figure 5.15.

The data were corrected in the time-domain for the wavelength dependence of the transit time of light in silica optical fiber, t_{fo}, using[75]

$$t_{fo} = \frac{l}{v}, \tag{5.1}$$

where l is the optical fiber length, and v is the group velocity of light in silica given by:

$$v = c\left[n(\lambda) - \lambda\frac{d(n(\lambda))}{d\lambda}\right]^{-1}, \tag{5.2}$$

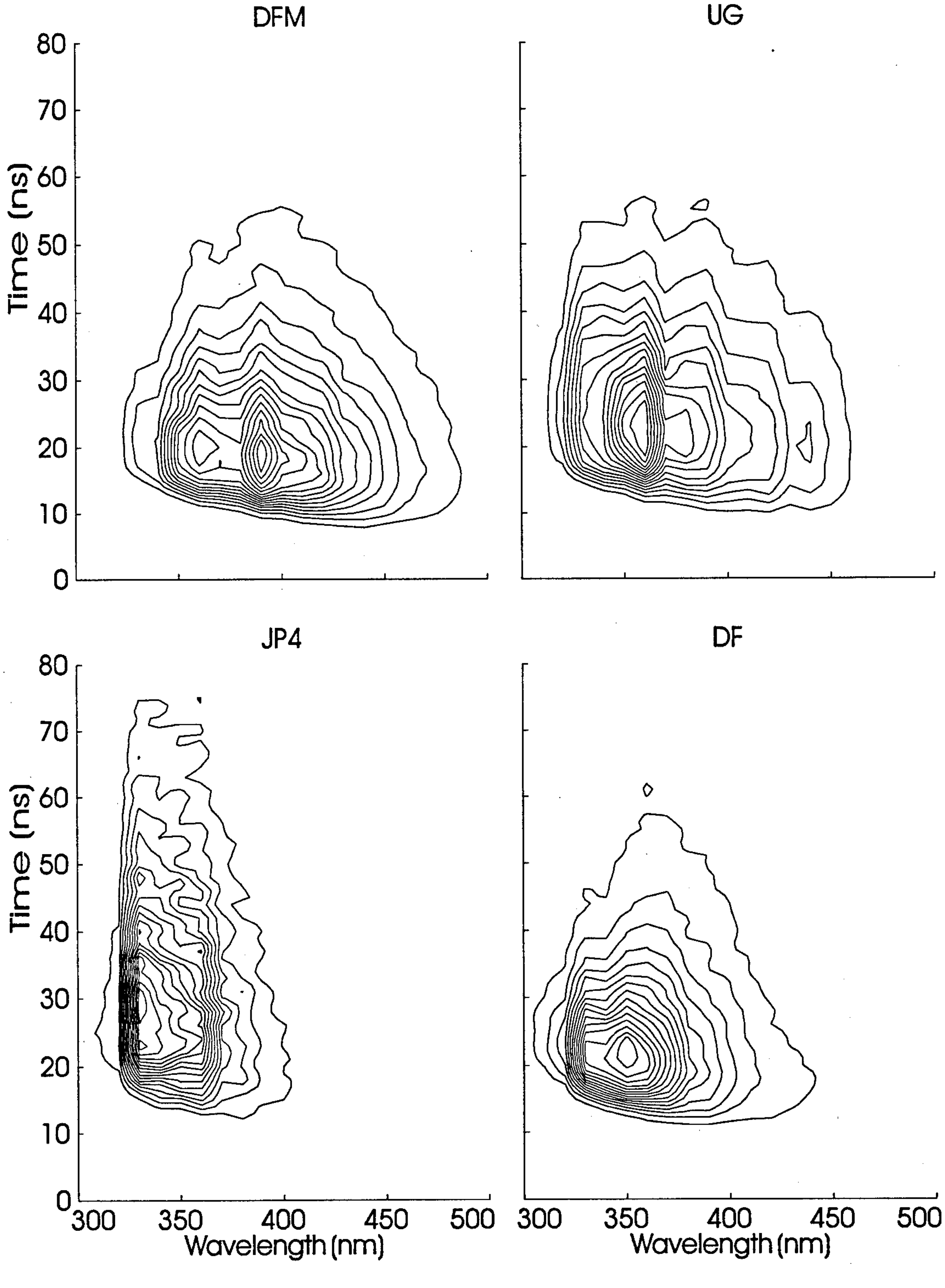


Figure 5.15. WTMs of four fuels on soil. Contours represent levels of equal fluorescence intensity. The weakest intensity is represented by the outermost contour in each WTM. Each WTM here has 15 contours, however, not all have the same intensity values. This effectively scales the WTMs for display purposes.

where c is the speed of light *in vacuo*, and $n(\lambda)$ is the index of refraction for silica. The index of refraction was determined with the dispersion equation:[76]

$$n(\lambda)=\left[1+\frac{c_1\lambda^2}{\lambda^2-c_2^2}+\frac{c_3\lambda^2}{\lambda^2-c_4^2}+\frac{c_5\lambda^2}{\lambda^2-c_6^2}\right]^{\frac{1}{2}}, \qquad (5.3)$$

where λ is wavelength, and the constants, c_i, are

$$\begin{aligned} c_1 &= 0.6961663, \\ c_2 &= 0.0684043, \\ c_3 &= 0.4079426, \\ c_4 &= 0.1162414, \\ c_5 &= 0.8974794, \\ \text{and } c_6 &= 9.86161. \end{aligned}$$

Data points are lost at the lower and upper time limits due to the correction. There were 107 time increments after correction. Figure 5.16 contains the same data as in Figure 5.15 after the correction for group velocity time shift. Note that the data in Figure 5.16 have a more "square" appearance than Figure 5.15.

After the evaluation, it was noted from the soil background measurements that several of the sapphire windows contained fluorescent impurities (*vide infra*), a conclusion that was verified by direct measurements on the windows themselves. A WTM contour plot for one of the fluorescent sapphire windows is displayed in Figure 5.17. Unfortunately, no record was kept of which of the 18 different sapphire windows was used for any particular measurements made during the week-long exercise.

The data were prepared for multimode analysis by arranging the 12 WTMs of the known fuel-soil mixtures, the 24 WTMs of the unknown fuel-soil mixtures, and four background WTMs (three with and one without window fluorescence) into a 3-array of

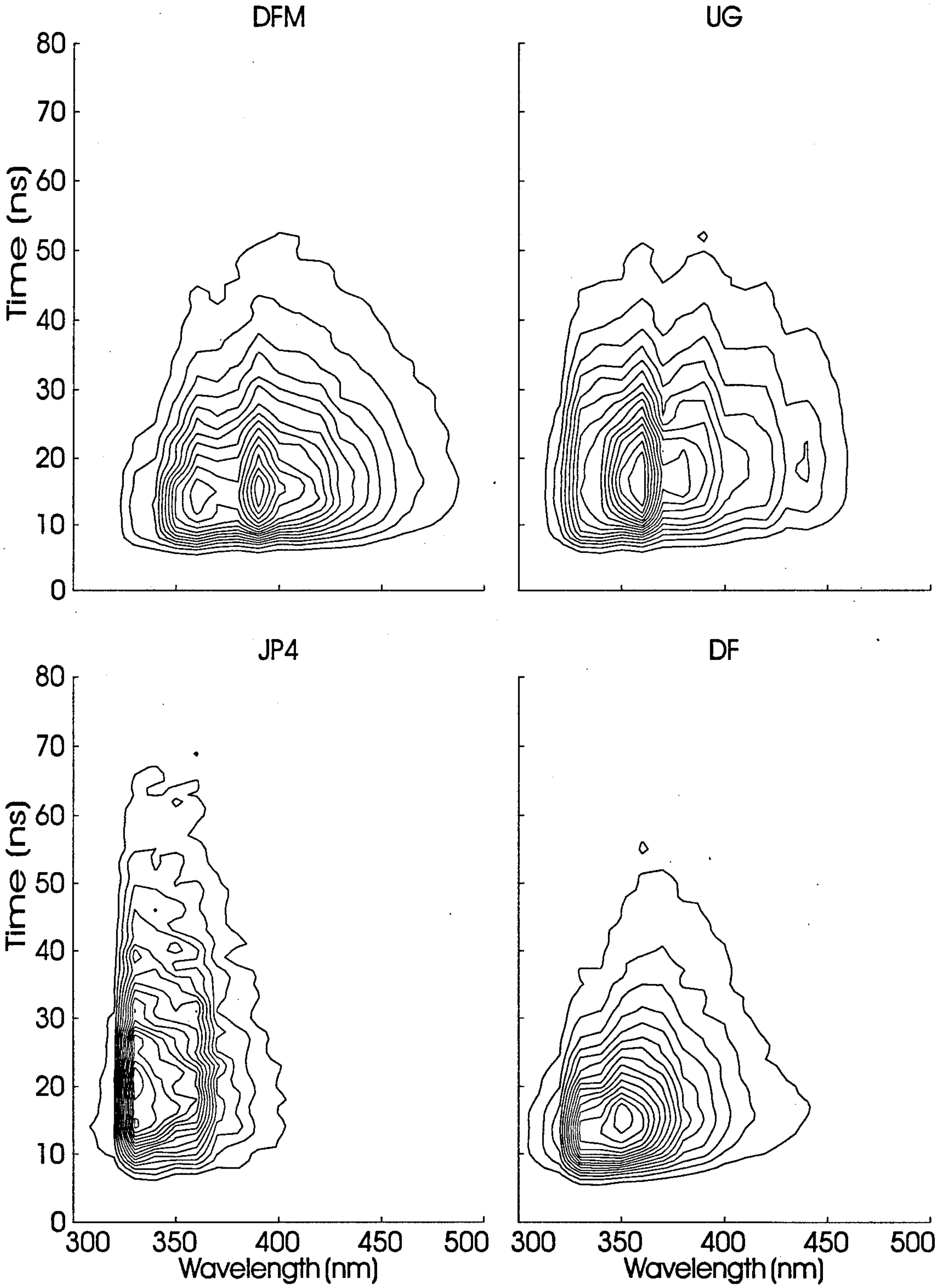


Figure 5.16. Time shift corrected WTMs of four fuels on soil. Contours represent levels of equal fluorescence intensity. The weakest intensity is represented by the outermost contour in each WTM. Each WTM here has 15 contours, however, not all have the same intensity values. This effectively scales the WTMs for display purposes.

dimension 21 by 107 by 40. An alternative analysis scheme, which will be discussed in Section 5.2.3, reduced the data order to a 2-array by integration along the time-mode. Integration was accomplished simply by summing along the time-mode of the WTMs. Since the time intervals are evenly spaced, this procedure is identical to a trapezoidal rule integration and results in a 21 by 40 matrix of time-integrated spectra.

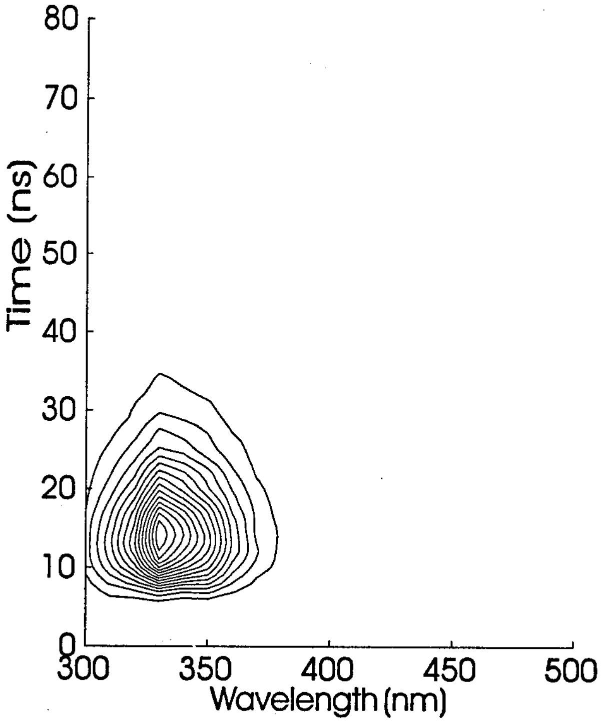


Figure 5.17. Background fluorescence including contaminated sapphire window.

5.2.1 Rank Estimation of Fuel Data

Rank estimation was carried out on the fuel fluorescence data in the same manner as applied to the two previous data sets. Rank was also estimated from the matrix of time-integrated spectra. Analyses were performed on the data both with and without the four background WTMs to evaluate their effect. They are presented in Table 5.4; results on the individual WTMs have been omitted for brevity, and the time-mode autocorrelation estimate was not pursued.

Table 5.4. Rank estimates for fuel fluorescence

Data Arrangement[1]	Number of WTMs[2]	IND[3]	*F*-test[4] (5 %)	$C(\mathbf{u}(\lambda))$[5]
Σ(WTMs)	36	7	4	3
$\int (\mathrm{WTM})\, dt$	36	15	5	3
Σ(WTMs)	40	6	4	3
$\int (\mathrm{WTM})\, dt$	40	11	4	4

[1] Data formed as matrix sum of WTMs or as a matrix of time-integrated spectra.
[2] Indicates whether only 36 fuel WTMs or 4 background WTMs were included.
[3] Malinowski's factor indicator.
[4] Malinowski's *F*-test for significance above the 5% level.
[5] Autocorrelation coefficient, values above +0.5.

Based on the results for the two previous data sets, it is tempting to concentrate on the rank estimates given by the *F*-test and the wavelength-mode autocorrelation. These indicate a rank of around four for this data.

The factor indicator function appears to be unstable with respect to the form of the data. This is probably a result of the change in the dimension of the data matrix, since the dimensions are used in the computation of the indicator values. The summed WTMs matrix is 21 by 107, while the time-integrated matrix is 21 by 40 or 21 by 36. The trend for the factor indicator function is to increase as the larger dimension is decreased. This may be a weakness of the factor indicator function. As Malinowski[46] pointed out, the factor indicator function is an empirical function and should be used with caution.

In light of the fact that the fuels are known to be complex mixtures with many different PAHs, the results presented in Table 5.4 are surprising. It is possible that there is only a small number of emitter groups and the individual compounds within the groups have very similar emission spectra and lifetimes. Other considerations include the level of experimental noise, species interactions, matrix effects, and group velocity correction errors for the fiber optics. All of these factors may contribute to the rank estimate being far lower than expected for these data.

5.2.2 TLD of Fuel Fluorescence

Mitchell and Burdick[77] have proposed classifying mixtures via trilinear decomposition of three-mode data. For example, consider a 3-array consisting of WTMs of many different fuel samples. Two of the modes from the trilinear decomposition would represent wavelength and fluorescence decay time, and the third mode would then represent a mixture or fuel-type mode. The relative proportions of the factors in the class-mode could be used as descriptors in a classification scheme, thereby exploiting the second-order advantage. One could not only classify the fuel, but also identify its major components.

To obtain a unique decomposition for 3-mode data, the matrices which form the underlying triple product of the data must meet the condition of Equation B.40, namely that the sum of the *k*-ranks of the matrices must be greater than or equal to two times the rank of the 3-array plus two. Thus, if even one of those matrices has two identical columns, a unique decomposition is not possible with 3M-ALS.

Is it likely that the fuel data collected have two identical or proportional columns in the factor matrices that make up the 3-array? Since the fuels are complex mixtures containing tens or even hundreds of fluorescent species and since there are only four fuels, it is likely that the mixture-mode factor matrix has *k*-rank of one. Therefore, we expect that the 3-array cannot be uniquely decomposed with 3M-ALS. Even considering the matrix effects of the different soils, which might contribute to variability among the fluorescence from different species, the number of fluorophores is likely too large to obtain a unique decomposition.

Another consideration in analyzing fuel emission is the possibility of nonlinear spectroscopic effects arising from interactions between species. Some of the potential sources of deviations from trilinearity for individual fluorophores are quenching, energy transfer, and reabsorption of fluorescence.

The more factors that are sought, the longer is the required computational time. Trilinear decomposition of this data set by 3M-ALS or 3M-NNALS took as long as a day (using a 120 MHz Pentium® computer and employing strict convergence criteria) for assumed rank of ten or greater, which is impractically long for many applications. Given these complications, what can be expected or gained from a trilinear decomposition here?

Eastwood[78] has suggested that the sources of fluorescence in complex mixtures, such as fuels, could be grouped into classes of fluorophores. The fluorescence spectral and lifetime data offer some encouragement. For example, single ring compounds typically emit in the short wavelength range, say 280 nm to 300 nm, and their lifetimes are also usually short in air-saturated aqueous solution. Naphthalene and its derivatives have intermediate wavelength emission (300 nm - 360 nm) and lifetimes (ca. 25-35 ns) emission. And the higher molecular weight PAHs tend to emit at even longer wavelengths; however, the trends in lifetime are less obvious for the PAHs. Perhaps this is the best that can be expected from fluorescence data of fuel products in view of the results from the simple solutions. Recall that in the four-component case of Data Set Two, we had to employ 3M-NNALS to obtain a satisfactory result.

If the goal of the decomposition is extraction of the fluorescence factors of all species contributing to the signal, trilinear decomposition is probably not a realistic approach; and more traditional analytical methods such as GC-MS or HPLC should be considered. However, if the goal is to find factors which adequately describe the fuels while taking advantage of the third-order nature of the data and generate good descriptors in the process, then 3M-NNALS may prove useful.

Given the previous rationalizations, we attempted to keep the rank of the decomposition as low as possible, but large enough to model the data and the sapphire window interference. Resolutions of ranks three through eight were obtained using EBP decomposition. All resulted in complex (imaginary) factors. The similarity transform method was employed to eliminate the complex factors. The resulting factors had generally large negative amplitude contributions with little, if any, physical significance. A factor clearly associated with the window fluorescence did not emerge from the analysis.

The EBP computations were followed by 3M-ALS. The results from the unrestricted 3M-ALS were also not very encouraging. The calculations tended to produce two or three very similar factors. Although these did not always meet the definition of a two-factor degeneracy, they were usually highly correlated in all three modes and were physically meaningless. A factor indicative of the sapphire window fluorescence could be identified in some resolutions, but not in others. Computation times were very long, as greater than 50,000 iterations were required to reach convergence in several cases.

3M-NNALS was performed for ranks three through seven using the results of the similarity transform method decomposition as starting vectors. The results from these calculations appeared more like real factors. However, rather than having smooth asymptotic transitions to the baseline, the mode features often tended to get zeroed out or “cut off” as they reached zero intensity or concentration for many factors. This is an unfortunate artifact of the non-negativity algorithm. On the other hand, the fact that none of the decompositions performed with this method required over 300 iterations for convergence is encouraging.

The rank five resolution by 3M-NNALS that produced the smallest residual is presented in Figure 5.18. Unfortunately, 3M-NNALS (at least in its current implementation) is very susceptible to trapping in local minimum solutions. At least two different resolutions were determined for the rank five case. Both resolutions produced reasonable representations of the window contribution, and the other factors were similar.

Three background samples which obviously contained window fluorescence were analyzed separately, also. The background WTMs with window fluorescence were trivial to analyze and proved to be rank one, based on the appearance of rank one through three 3M-NNALS resolutions. The factor for the interference is displayed for comparison in the last frame of Figure 5.18.

The rank five resolution was chosen from among all of the decompositions because it provided the best match to the window factors obtained from the 3M-NNALS analysis of

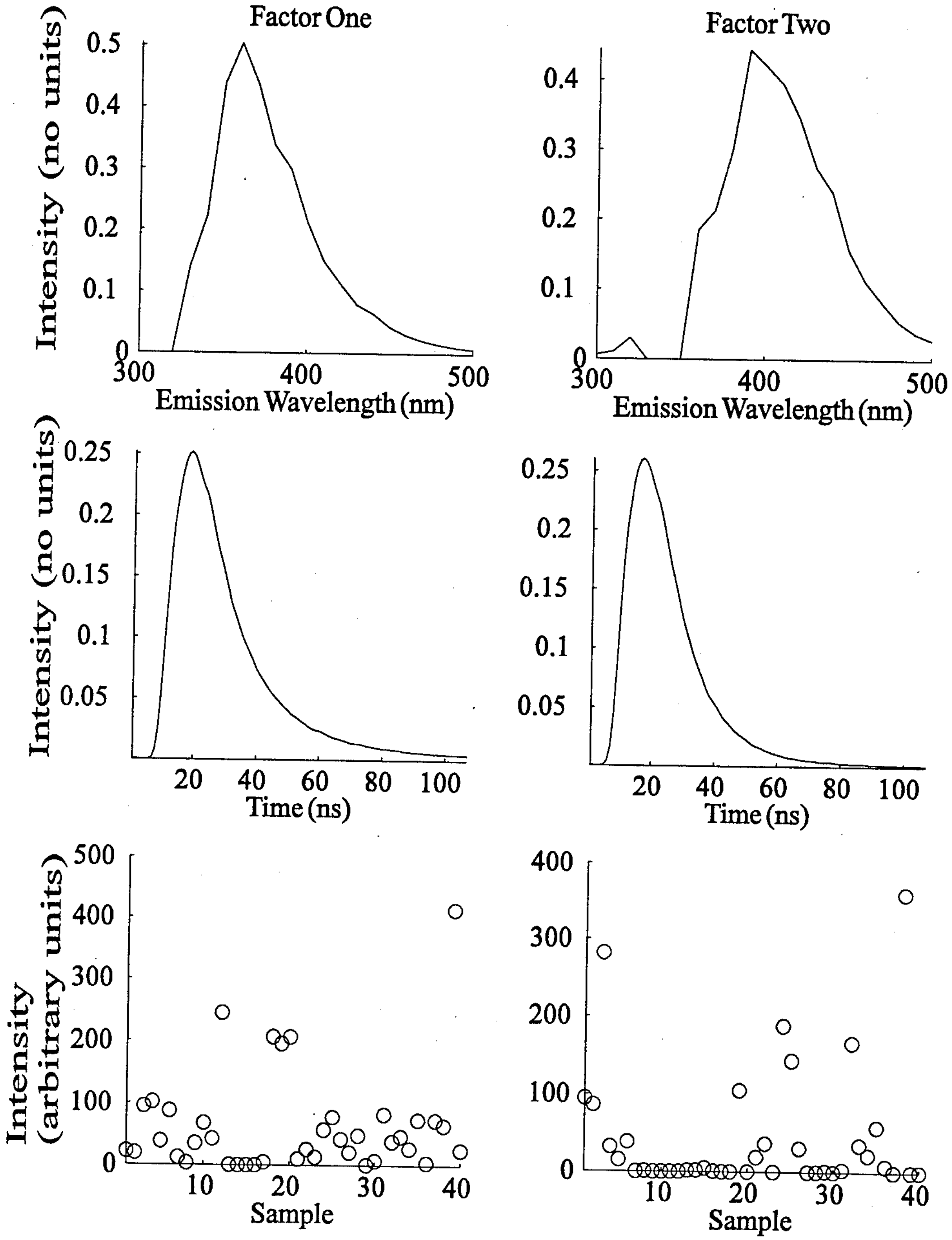


Figure 5.18. Rank five 3M-NNALS decomposition factors for fuel fluorescence.

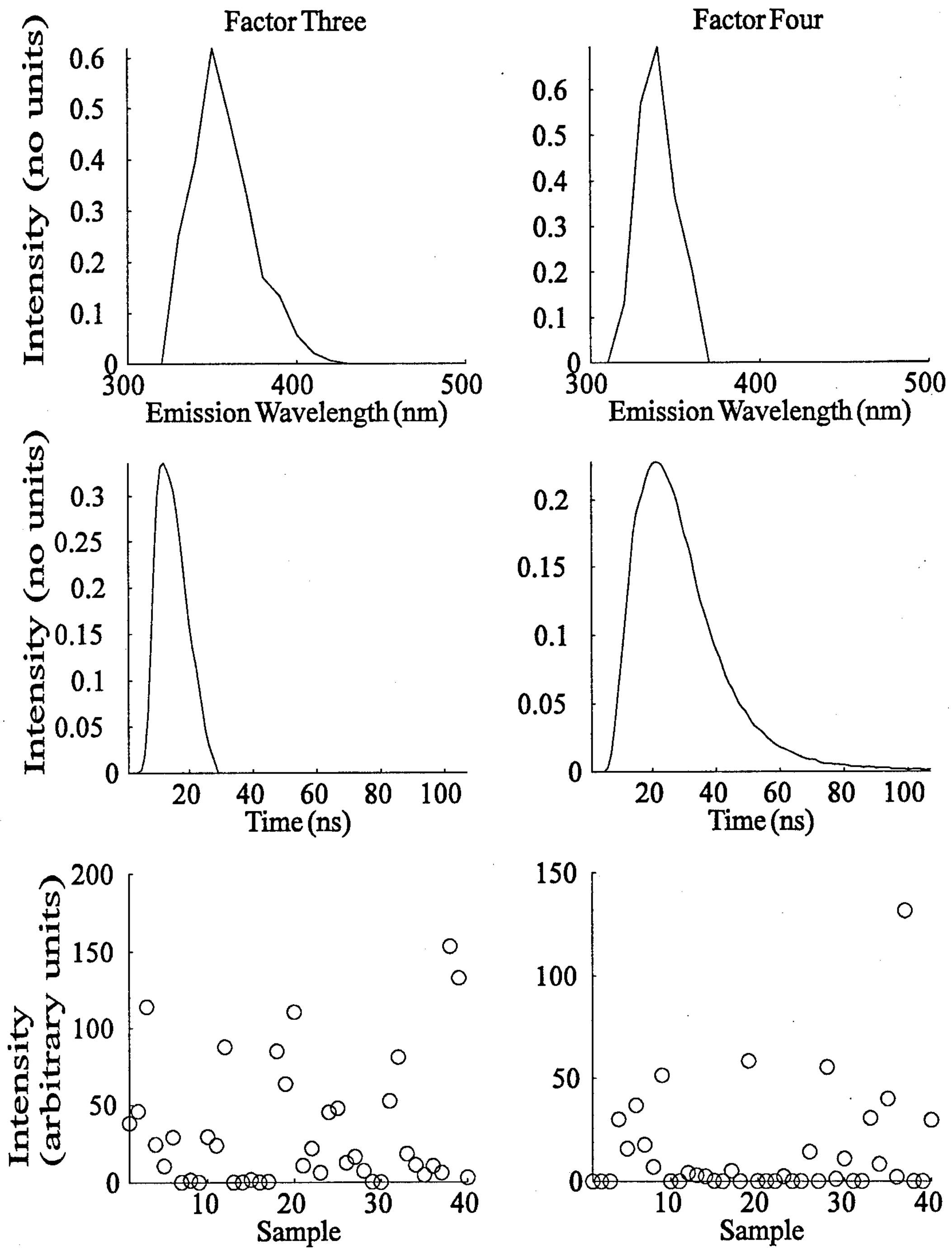


Figure 5.18. Continued.

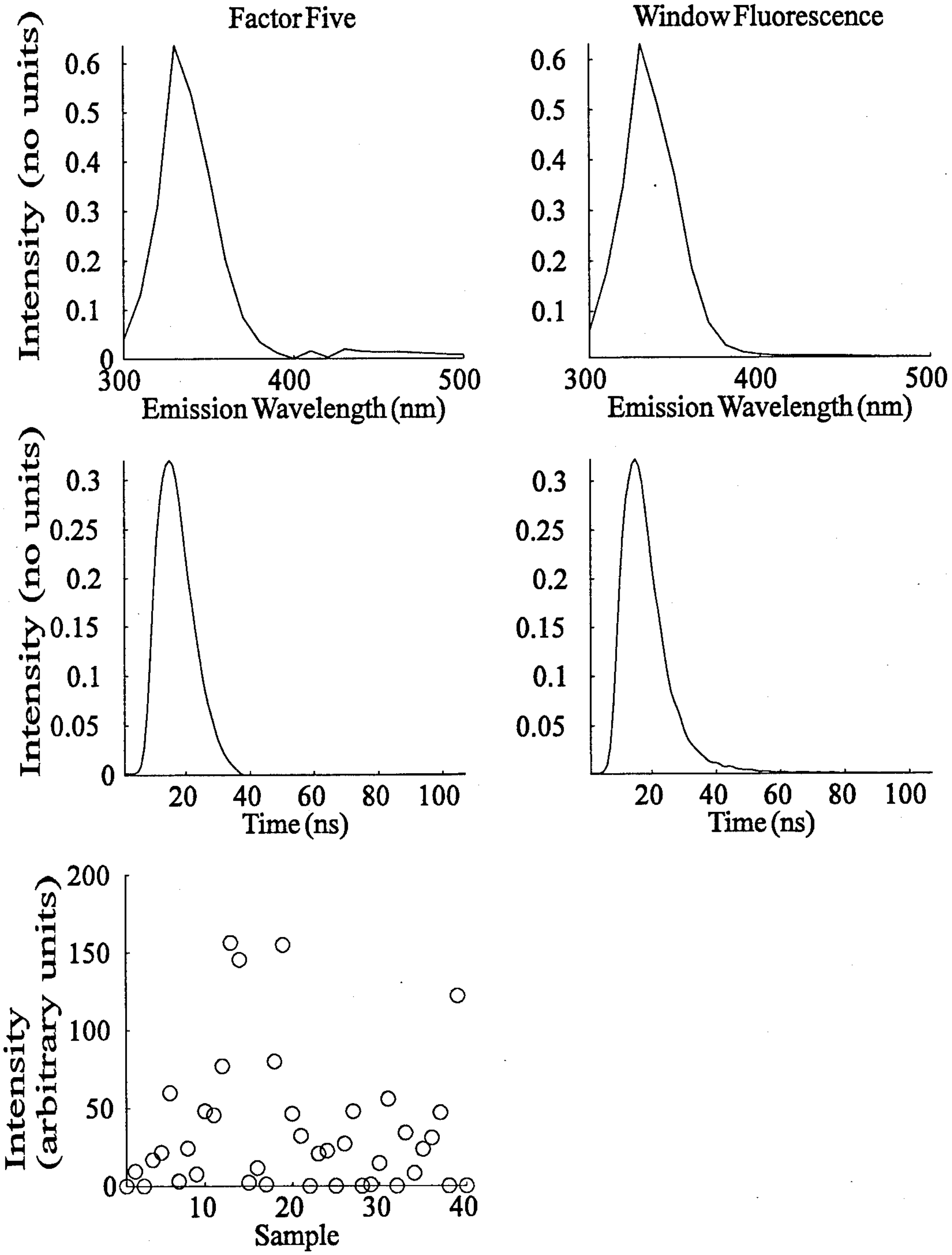


Figure 5.18. Continued.

just the background WTMs themselves. Clearly, the fifth factor of the rank five resolution represents the interference. It highly correlates with the background resolution in both the wavelength and time modes; the angle cosines of the overlap are 0.9973 and 0.9984 in the respective modes. The results for the higher rank resolutions are similar to those of rank five. The rank four resolution did not provide a clear identification of the window factor.

It is unfortunate that the features appear to get "cut off" for a number of the factors. Nevertheless, the window fluorescence can be identified and removed from each sample WTM by subtracting the triple product of the vectors for the three modes of Factor Five in Figure 5.18 from the 3-array.

Sample-mode Factors One through Four were used as descriptors in a classification scheme (see Section 5.2.3).

5.2.3 Analysis of Fuel Fluorescence With Self-modeling Curve Resolution

A fundamental question about the multimode data generated by a ROST™ or similar system is whether all this information is really needed. Certainly first-order data, in the form of spectra, can be collected very efficiently using a charge coupled device (CCD) or photo diode array (PDA) detector. The time-integrated data created during rank estimation represents such first-order data. Such data can be utilized to compare the merits of higher order data with lower order data.

The time-integrated data were subjected to a self-modeling curve resolution (SMCR) scheme using a two-mode NNALS algorithm. Decompositions were obtained for ranks three through seven. The wavelength-mode results, in many cases, were similar to those of the wavelength-mode from 3M-NNALS. However, we found that the extracted factors were very sensitive to starting points. An important aspect of the analysis was extracting the sapphire window component. In many cases of profile extraction here, a single window factor did not emerge from the calculation. This was the case for our rank five minimum residual result.

5.2.4 Comments on TLD of Fuel WTMs

3M-NNALS is the only feasible way to analyze WTMs from complex mixtures such as fuels. EBP and 3M-ALS were not able to obtain physically meaningful factors from these data, and the latter was far too time intensive.

Assignments of the factors to groups of compounds, such as naphthalenes and substituted naphthalenes, etc., is premature; and further analysis of fuel fluorescence in needed. Performing 3M-NNALS on simulated fuel mixtures which are chemically and spectroscopically similar to fuels, but of known composition, may aid in this process. Rank estimation performance should also benefit from this type of research.

The true significance of the TLD technique is its ability to identify and remove interferences from the data and generate descriptors of the fuels. 3M-NNALS was able to isolate a factor closely resembling the sapphire window interference in the rank five resolution, an accomplishment that two-mode analysis could not match. How successfully and completely the technique removed the window factor cannot be characterized for these data because the true amount of window interference is not known. But in the analysis that follows, we can speculate on its performance based on classification results with and without the window component. More work needs to be performed in this area also.

5.3 Classification of Fuels Using Fluorescence Data

One element of the field laser performance evaluation was to determine the ability of the instruments to classify the 24 unknown fuel-soil mixtures using the 12 known fuel-soil mixtures as a training set.

To separate the fuel groups, we used Fisher's discriminant function, described in Section 4.1.2, to generate linear discriminants with which to classify the unknowns. Multivariate normal distributions and equal covariance matrices for the four fuels were assumed. Prior probabilities were equal under the conditions of the evaluation. Although the classification algorithm presented in Section 4.1.2 is reasonably straightforward, the challenge is to find the most accurate descriptors of the fuels within the data.

In addition to using the mixture-mode factors, descriptors for the classification algorithm were obtained in several other ways. One method used the mixture-mode elements from three-mode. Another method used the mixture-mode PFA factors (computed using SVD) of the time-integrated spectra as descriptors. Each of the methods used was evaluated on its ability to predict fuel type of the 24 unknowns. Normalization of the descriptor vectors was performed in all cases to minimize the effects of concentration and the soil matrix. The normalization consisted of scaling the vector comprising the descriptors of each fuel to unit length. However, the stage of the classification process at which normalization was performed depended on how the descriptors were obtained.

5.3.1 3M-NNALS Mixture-mode Factors as Descriptors

The elements of a 3M-NNALS mixture-mode factor represent the contribution of the corresponding wavelength-time dyad to the WTMs. Thus, the third element of mixture-mode Factor One in Figure 5.18, for example, is the contribution of the wavelength-time dyad of Factor One needed to model the WTM of the third mixture (which happens to be a DFM on sand combination). Each element of each mixture-mode factor is a descriptor of one of the fuels (or window). A rank five decomposition yields five descriptors for each fuel; a rank six decomposition, six; and so forth.

Table 5.5 contains the classification results utilizing all of the factors, including the window factors, from 3M-NNALS decomposition for ranks four through seven.

To eliminate the window fluorescence from the data, the obvious window factor was identified for each rank decomposition, and its mixture-mode factor was omitted as a descriptor. The descriptor vectors were normalized and used in classification. The classification results with the window contributions omitted are in Table 5.6.

The classification rates were good for all rank decompositions, with and without the window factors. In all cases of misclassification, the mixtures corresponded to the lowest fuel concentration (1000 ppm) in the evaluation samples. In fact, the most consistently misclassified mixture, Unknown M, was a low concentration of JP4, the weakest emitter of the fuels, and was mixed on the CAFB soil, which was the poorest soil matrix in terms of fluorescence response. Support for this last statement comes from the fact that 10 of the 12 misclassifications in Table 5.5 occurred for CAFB soil samples; in Table 5.6, eight of the 10 misclassifications occurred on the CAFB soil. Unknown M had a very weak emission and the poorest S/N of any sample in the study.

We suspect that the second most misclassified mixture, Unknown T, has a very large sapphire window interference in proportion to its total fuel emission. This is also the case

Table 5.5. Classification of unknown fuels using 3M-NNALS factors

Unknown[1]	Four Factors	Five Factors	Six Factors	Seven Factors	Actual Fuel	Conc. ppm ($\times 10^{-3}$)	Soil Type
A	JP4	JP4	JP4	JP4	JP4	1	CLNAS
B	DF	DF	DF	DF	DF	10	CLNAS
C	UG	UG	UG	UG	UG	10	Sand
D	DF	DF	DF	DF	DF	10	CAFB
E	UG	UG	UG	DFM	DFM	1	CAFB
F	DFM	DFM	DFM	DFM	DFM	1	CLNAS
G	DF	DF	DF	DF	DF	1	CAFB
H	DFM	DFM	DFM	DFM	DFM	10	CLNAS
I	DFM	DFM	DFM	DFM	DFM	10	CAFB
J	UG	UG	UG	UG	UG	1	Sand
K	DF	DF	DF	DF	DF	1	CLNAS
L	JP4	JP4	JP4	JP4	JP4	10	CLNAS
M	UG	UG	UG	UG	JP4	1	CAFB
N	JP4	JP4	JP4	JP4	JP4	1	Sand
O	DF	DF	DF	DF	DF	1	Sand
P	DFM	DFM	DFM	DFM	DFM	1	Sand
Q	UG	UG	UG	UG	UG	10	CAFB
R	DFM	UG	DFM	UG	UG	1	CLNAS
S	UG	UG	UG	UG	UG	10	CLNAS
T	DF	UG	DF	DF	UG	1	CAFB
U	JP4	JP4	JP4	JP4	JP4	10	Sand
V	DFM	DFM	DFM	DFM	DFM	10	Sand
W	DF	DF	DF	DF	DF	10	Sand
X	JP4	JP4	JP4	JP4	JP4	10	CAFB
Total Correct	20 83%	22 92%	20 83%	22 92%			

[1] Unknowns A-X correspond to Mixtures 17-40 in Figure 5.18.

Table 5.6. Classification of unknown fuels using 3M-NNALS factors omitting suspected window factor

Unknown[1]	Four	Five	Six	Seven	Actual	Conc.	Soil
	Factors[2]	Factors[2]	Factors[2]	Factors[2]	Fuel	ppm ($\times10^{-3}$)	Type
A	JP4	JP4	JP4	JP4	JP4	1	CLNAS
B	DF	DF	DF	DF	DF	10	CLNAS
C	UG	UG	UG	UG	UG	10	Sand
D	DF	DF	DF	DF	DF	10	CAFB
E	DFM	DFM	UG	DFM	DFM	1	CAFB
F	DFM	DFM	DFM	DFM	DFM	1	CLNAS
G	DF	DF	DF	DF	DF	1	CAFB
H	DFM	DFM	DFM	DFM	DFM	10	CLNAS
I	DFM	DFM	DFM	DFM	DFM	10	CAFB
J	UG	UG	UG	UG	UG	1	Sand
K	JP4	DF	DF	DF	DF	1	CLNAS
L	JP4	JP4	JP4	JP4	JP4	10	CLNAS
M	UG	UG	UG	UG	JP4	1	CAFB
N	DF	JP4	JP4	JP4	JP4	1	Sand
O	DF	DF	DF	DF	DF	1	Sand
P	DFM	DFM	DFM	DFM	DFM	1	Sand
Q	UG	UG	UG	UG	UG	10	CAFB
R	UG	UG	UG	UG	UG	1	CLNAS
S	UG	UG	UG	UG	UG	10	CLNAS
T	UG	DF	DF	DF	UG	1	CAFB
U	JP4	JP4	JP4	JP4	JP4	10	Sand
V	DFM	DFM	DFM	DFM	DFM	10	Sand
W	DF	DF	DF	DF	DF	10	Sand
X	JP4	JP4	JP4	JP4	JP4	10	CAFB
Total	21	22	21	22			
Correct	88%	92%	88%	92%			

[1] Unknowns A-X correspond to Mixtures 17-40 in Figure 5.18

[2] The number of factors used in classification is one less indicated since the window component was removed.

for Unknown E, although to a lesser degree.

The integrated fluorescence intensities versus unknown mixture are displayed in Figure 5.19. The estimated window fluorescence has been removed from the fluorescence before integration using the rank five 3M-NNALS window factor. Included in Figure 5.19 are the integrated window fluorescence estimates for each mixture. Note the intensities and window components of the misclassified mixtures. It is obvious here which unknowns suffer from the greatest interference.

In general, fuel fluorescence intensity follows the trend: DFM is slightly stronger than DF, which is stronger than UG, which is much stronger than JP4. The trend for soil effects based on fuel fluorescence intensity on the respective soil is fuel-sand intensity is much greater than fuel-CLNAS soil mixture intensity, which is greater than fuel-CAFB soil intensity.

It is surprising and disappointing that the number of misclassifications does not improve more when the suspected window component is omitted from the set of descriptors. A disturbing aspect of this analysis is that three classifications, Unknowns K and N in the Four Factor column and Unknown T in the Five Factors column, went from correct with the window descriptor included to incorrect when it was omitted. It is also difficult to establish a pattern for misclassification in the presence of window fluorescence. Window fluorescence may make UG appear as DFM, as in the case of Unknown R, or the reverse, as in the case of Unknown E.

Apparently, the two conditions, weak emission and large proportional window fluorescence contribution to total signal, are the most significant causes of misclassification. However, it is not clear yet that the results of Table 5.6 are the best which can be achieved.

Conceivably, the trilinear model employed here has failed to model subtleties in the data which are important in classification. This may be a fault with using the 3M-NNALS factors as descriptors, since they are not the true least squares estimates of the data. It is possible that PFA, with its orthonormal basis set spanning the true factor space, may be able to generate better descriptors for the weaker intensity fuel mixtures.

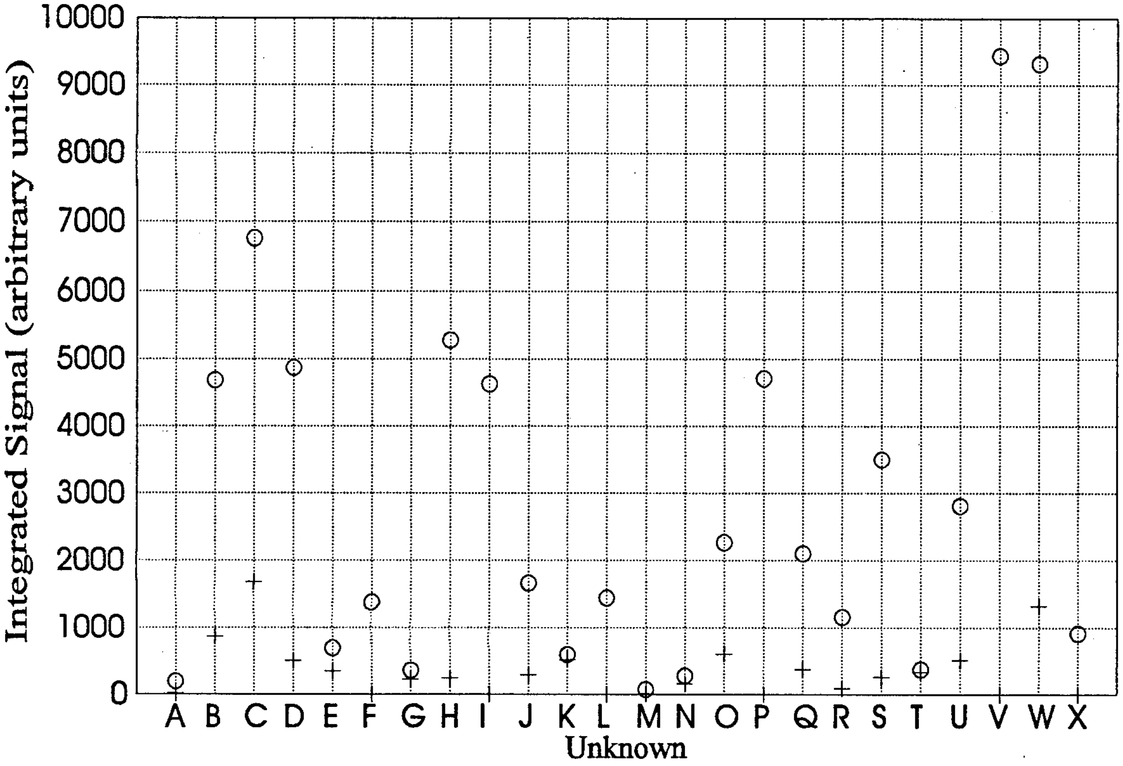


Figure 5.19. Unknown fuel fluorescence and sapphire window fluorescence. Integrated fluorescence intensities for unknown mixtures are represented by o's. Integrated window fluorescence intensities are represented by + signs.

5.3.2 Three-mode PFA Mixture-mode Factors as Descriptors

The rank five sapphire window factor, Factor Five of Figure 5.18, was subtracted from the WTMs of the 12 training mixtures and 24 unknowns. The resulting 3-array was analyzed using the TUCKER1 (eigenanalysis on the unfolded 3-array without performing ALS), three-mode PFA.

The mixture-mode eigenvectors were multiplied by their corresponding eigenvalues and used as descriptors. Before discrimination and classification, the descriptors were normalized as in the Section 5.3.1. The results for various numbers of PFs are presented in Table 5.7.

The results of the classification here are excellent. One interpretation that can be made is that the factors from the 3M-NNALS do not adequately represent the data. This is probably a result of substantial deviations from trilinearity of the fuel WTMs. However, the trilinear model seems to be adequate to fit and extract the sapphire window fluorescence.

The TUCKALS3 algorithm was also employed to generate descriptors with identical results.

5.3.3 Principal Factors of Time-integrated Spectra as Descriptors

The WTMs of the 12 training mixtures and the 24 unknowns were time integrated both before and after correction for group velocity shift of the fiber optics. The data were arranged into a 36 by 21 matrix and subjected to PFA using SVD such that the left singular vectors spanned the row space (i.e., the mixture-mode space) of the matrix and the right singular vectors spanned the column space (i.e., the wavelength-mode space) of the matrix.

Table 5.7. Classification of unknown fuels using three-mode PFA mixture-mode PFs

Unknown	3 PFs[1]	4 PFs[1]	5 PFs[1]	6 PFs[1]	Actual Fuel
A	JP4	JP4	JP4	JP4	JP4
B	DF	DF	DF	DF	DF
C	UG	UG	UG	UG	UG
D	DF	DF	DF	DF	DF
E	DFM	DFM	DFM	DFM	DFM
F	DFM	DFM	DFM	DFM	DFM
G	DF	DF	DF	DF	DF
H	DFM	DFM	DFM	DFM	DFM
I	DFM	DFM	DFM	DFM	DFM
J	UG	UG	UG	UG	UG
K	DF	DF	DF	DF	DF
L	JP4	JP4	JP4	JP4	JP4
M	JP4	JP4	JP4	JP4	JP4
N	JP4	JP4	JP4	JP4	JP4
O	DF	DF	DF	DF	DF
P	DFM	DFM	DFM	DFM	DFM
Q	UG	UG	UG	UG	UG
R	UG	UG	UG	UG	UG
S	UG	UG	UG	JP4	UG
T	UG	UG	UG	UG	UG
U	JP4	JP4	JP4	JP4	JP4
V	DFM	DFM	DFM	DFM	DFM
W	DF	DF	DF	DF	DF
X	JP4	JP4	JP4	JP4	JP4
Total Correct	24 100%	24 100%	24 100%	23 96%	

[1] Used rank five 3M-NNALS decomposition to remove sapphire window (interference) factor.

Descriptors were generated by taking the product of the left singular vector matrix with the singular value (diagonal) matrix. Then, they were entered into the linear discrimination and classification algorithm. The capsule results of the analyses are contained in Table 5.8.

Table 5.8. Number of correct fuel classifications with PFA of time-integrated spectra

Spectra Source[1]	3 PFs	4 PFs	5 PFs	6 PFs
Raw WTMs	20	21	22	21
Corrected WTMs	20	21	23	23

[1] Spectra generated from summing (time-integrating) WTMs before correcting for group velocity shift of fiber optics (Raw WTMs) and after correction (Corrected WTMs).

The misclassified fuels from these analyses were all low concentration. The same fuels were misclassified in the three PF and four PF cases. This was not the case with the five PF and six PF cases. Table 5.8 is not intended to illustrate a problem with time integration of spectra or the use of PFA on spectra or the group velocity shift correction of the data. Indeed, the spectra in the two cases differ only minutely due to losses of information on the edges of the time-mode. The PFs also have only minor differences even up to rank six.

Table 5.8 points out a possible problem with the classification algorithm. When the sample size of the standards (training set) is small compared to the number of descriptors, the covariance matrix estimates, $\mathbf{S}_i$ of equation 4.8, can become highly variable (and unstable upon inversion) and result in poor classification. The number of standards per class here is three. When the number of descriptors exceeds three, the classification may be suffering. The results in the PF 5 and PF 6 columns of Table 5.8 may be a manifestation of this phenomenon.

The discrepancy illustrated in Table 5.8 might have another cause. Recall that the rank estimation for the matrix of integrated spectra was four or five. Eigenvectors greater than four or five may be noise eigenvectors and disrupt the classification. Although the variance accommodated by the higher eigenvalues is very small, their effect may be more prominent on the weaker fluorescing mixtures.

After removing the window factors as in Section 5.3.2, the WTMs were time integrated and factor analyzed with SVD. The significant factors were then used in the discrimination and classification algorithms to classify the unknowns. Table 5.9 contains the classifications based on the 3M-NNALS, rank five window removal.

The results in Table 5.9 are superb. Apparently choosing to integrate and using orthogonal factors in the classification scheme was an excellent choice. Since these results are perfect, the inference can be made that classification of these fuels does not require three-mode data. However, the analyses required to get to this point speak for themselves concerning the three-mode nature of the data.

Subtraction of 3M-NNALS window factors other than the rank five decomposition were also performed. The classification results using PFs from the time-integrated spectra in these cases were not as good as the rank five window removal case. However, they were as good or better than the 3M-NNALS mixture-mode results presented in Table 5.6. Again, deviations from trilinearity in the data give advantages to the orthogonal factors to represent the data. Clearly though, the choice of window-mode factor is very important; thus, so is the choice of number of factors in the 3M-NNALS model.

Table 5.9. Classification of unknown fuels using PFA of time-integrated spectra

Unknown	3 PFs[1]	4 PF s[1]	5 PF s[1]	6 PF s[1]	Actual Fuel
A	JP4	JP4	JP4	JP4	JP4
B	DF	DF	DF	DF	DF
C	UG	UG	UG	UG	UG
D	DF	DF	DF	DF	DF
E	DFM	DFM	DFM	DFM	DFM
F	DFM	DFM	DFM	DFM	DFM
G	DF	DF	DF	DF	DF
H	DFM	DFM	DFM	DFM	DFM
I	DFM	DFM	DFM	DFM	DFM
J	UG	UG	UG	UG	UG
K	DF	DF	DF	DF	DF
L	JP4	JP4	JP4	JP4	JP4
M	JP4	JP4	JP4	JP4	JP4
N	JP4	JP4	JP4	JP4	JP4
O	DF	DF	DF	DF	DF
P	DFM	DFM	DFM	DFM	DFM
Q	UG	UG	UG	UG	UG
R	UG	UG	UG	UG	UG
S	UG	UG	UG	UG	UG
T	UG	UG	UG	UG	UG
U	JP4	JP4	JP4	JP4	JP4
V	DFM	DFM	DFM	DFM	DFM
W	DF	DF	DF	DF	DF
X	JP4	JP4	JP4	JP4	JP4
Total Correct	24 100%	24 100%	24 100%	24 100%	

[1] Used rank five 3M-NNALS decomposition to obtain sapphire window (interference) factor.

Analogous classification was performed on the two-mode data after subtraction of the window factors generated with SMCR. The results were an improvement over those in Table 5.8, but, in general, not as strong as in Table 5.9. Subtraction of the rank four SMCR window factor resulted in one perfect classification using four PFs.

5.3.4 Comments on Fuel Classification

This data set was particularly difficult to deal with because of the interference. Certainly, greater care could have been exercised in the selection of the sapphire windows used for the measurement. On the other hand, this type of problem is probably highly representative of those that will be encountered during *in situ* analysis, whether at a hazardous waste site or in human tissue. The effort expended here should facilitate analyses on actual unknowns.

Correct classification of the fuels was possible after removing the interference and modeling the data with orthogonal factors to generate descriptors. Major considerations of classification were the method of profile extraction, selection of window factor, choice of descriptors, and the limited size of the training set.

Based on these results and the results presented in Section 5.2, we believe that the best approach to profile extraction of fuel fluorescence is 3M-NNALS. While 3M-NNALS is susceptible to MLO, it is less so than SMCR.

Future studies should be conducted to include larger training sets with more fuel types. They should include various levels of oxygen to examine the effects of fluorescence quenching on profile extraction and classification.

5.4 Cluster Analysis

Cluster analysis could be extremely useful in distinguishing different types of luminescent materials, such as fuels or tissue types, without prior knowledge of types of materials present. Such a scheme could be applied to a contaminated site containing multiple fuels or to a collection of tissue specimens containing healthy and diseased tissues or tissue from different species.

The goal of this section is to evaluate the viability of HCA to perform clustering of fuels. In addition, we would like to find the best measure of similarity to be employed in the HCA algorithm.

As in most of the classification schemes used in this work, HCA was performed on data corrected for the sapphire window contribution by subtracting the window factor from the 3M-NNALS rank five decomposition. The HCA algorithm described in Section 4.2 was performed with Matlab® using a program written in this laboratory. Numerous approaches were taken for both generating of the similarity (dissimilarity) matrices and forming linkages during clustering.

Clustering was performed using all of mixture-mode PFs employed in the classifications of Section 5.3 as descriptors of the fuel data. Both "city block" and Euclidean metrics were tested as dissimilarity measures. Single, complete, and average linkage methods were investigated for forming linkages. All of these descriptors, the 3M-NNALS factors, PFs from time-integrated spectra, and PFs from three-mode PFA, performed clustering very well.

The average linkage dendrogram of the descriptor-vector-normalized 3M-NNALS factors, excluding the sapphire window factor, is displayed in Figure 5.20. The Euclidean metric was used to generate the distance matrix. Using a distance of around 0.4 as a group separation cutoff, only Unknowns T and M are incorrectly grouped. In fact, they do not

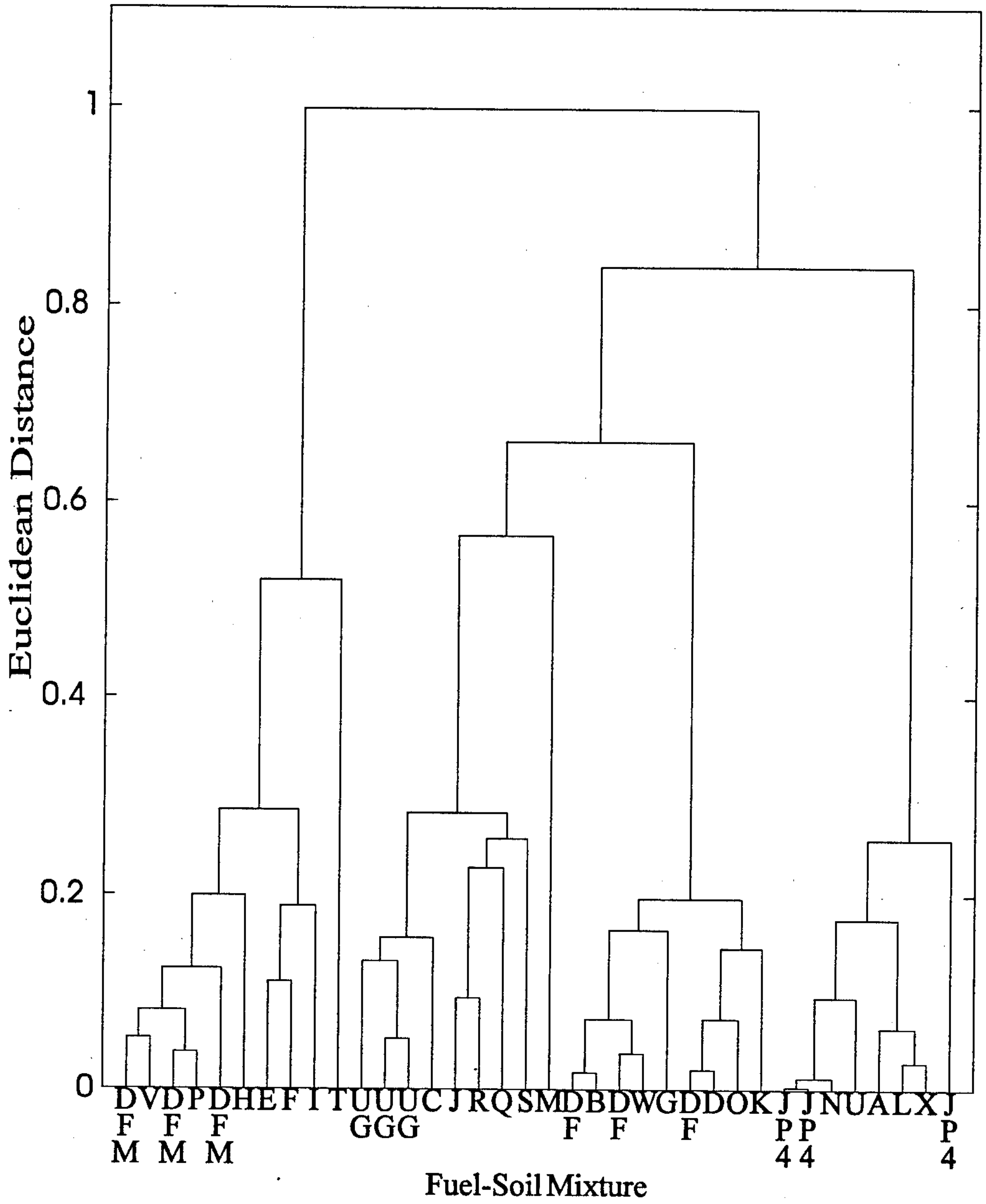


Figure 5.20. Average linkage dendrogram of fuels using 3M-NNALS factors. Normalized mixture-mode factors from rank five 3M-NNALS decomposition of fuels with sapphire window (interference) factor omitted. Similarity measure is Euclidean distance between descriptor vectors. The known spectra are indicated by their fuel abbreviations and the unknowns by their letter designators.

group with any of the major fuel groups. This is not too surprising given that these same two unknowns were misclassified using these descriptors (see Table 5.6).

Figure 5.21 portrays the average linkage dendrogram of the fuels using the first four left singular vectors of the 36 by 21 matrix of time-integrated spectra multiplied by their respective eigenvectors. These are the same data that were used in the 4 PF column of Table 5.9. With a distance cutoff of 0.2, only Unknown M is not grouped with any major group. Otherwise, the groupings are excellent.

PFs from three-mode PFA, used in 4 PF column of Table 5.7, were used to produce the average linkage dendrogram in Figure 5.22. Here the groupings are more ambiguous, although still very good. If one chooses a distance cutoff of 0.5, one of the JP4 standards and Unknown A are excluded from the JP4 grouping. Choosing a cutoff of 1.0 lumps the UG and DF fuels into a single group. Unknown M is incorrectly grouped into the UG cluster in any event.

When dealing with spectral vectors, the angle cosine is one of the more intuitive measures of similarity because it indicates the proximity of two vectors in space and is therefore a measure of correlation. The angle cosine calculation has normalization built into it; only the direction of the vector is considered, and the magnitude is unimportant. For these reasons, the remainder of this section will deal with clustering in which the angle cosine (or its value subtracted from unity) is taken as the measure of similarity.

Figure 5.23 contains the single linkage dendrogram using one minus the angle cosine of the spectra as the measure of similarity. There is excellent separation of the groups based upon these criteria. Only one mixture, Unknown M, did not group as expected, which is

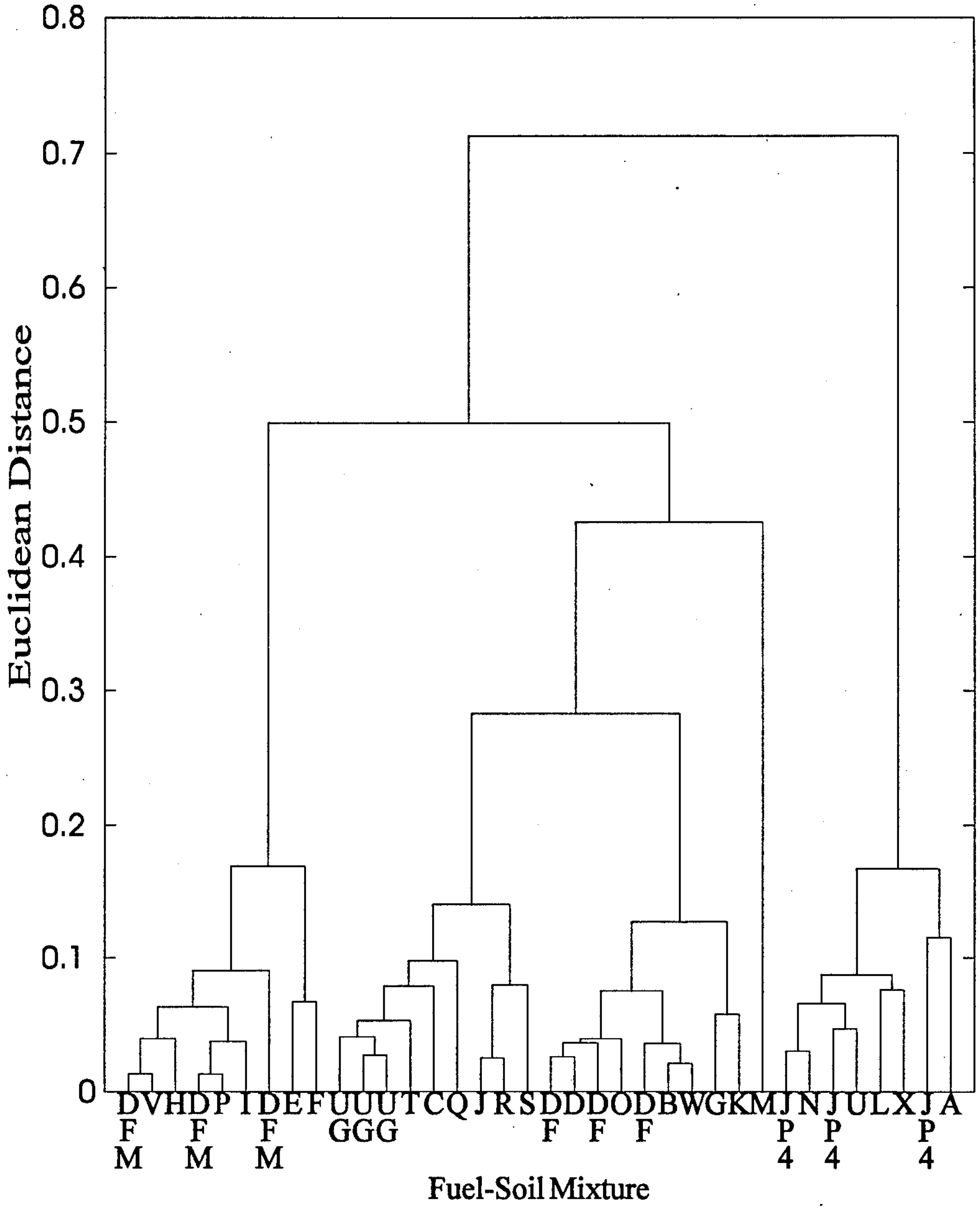


Figure 5.21. Average linkage dendrogram of fuels using PFA of time-integrated spectra. First four mixture-mode PFs of normalized spectra. Sapphire window (interference) factor removed using rank five 3M-NNALS. Similarity measure is Euclidean distance between descriptor vectors. The known spectra are indicated by their fuel abbreviations and the unknowns by their letter designators.

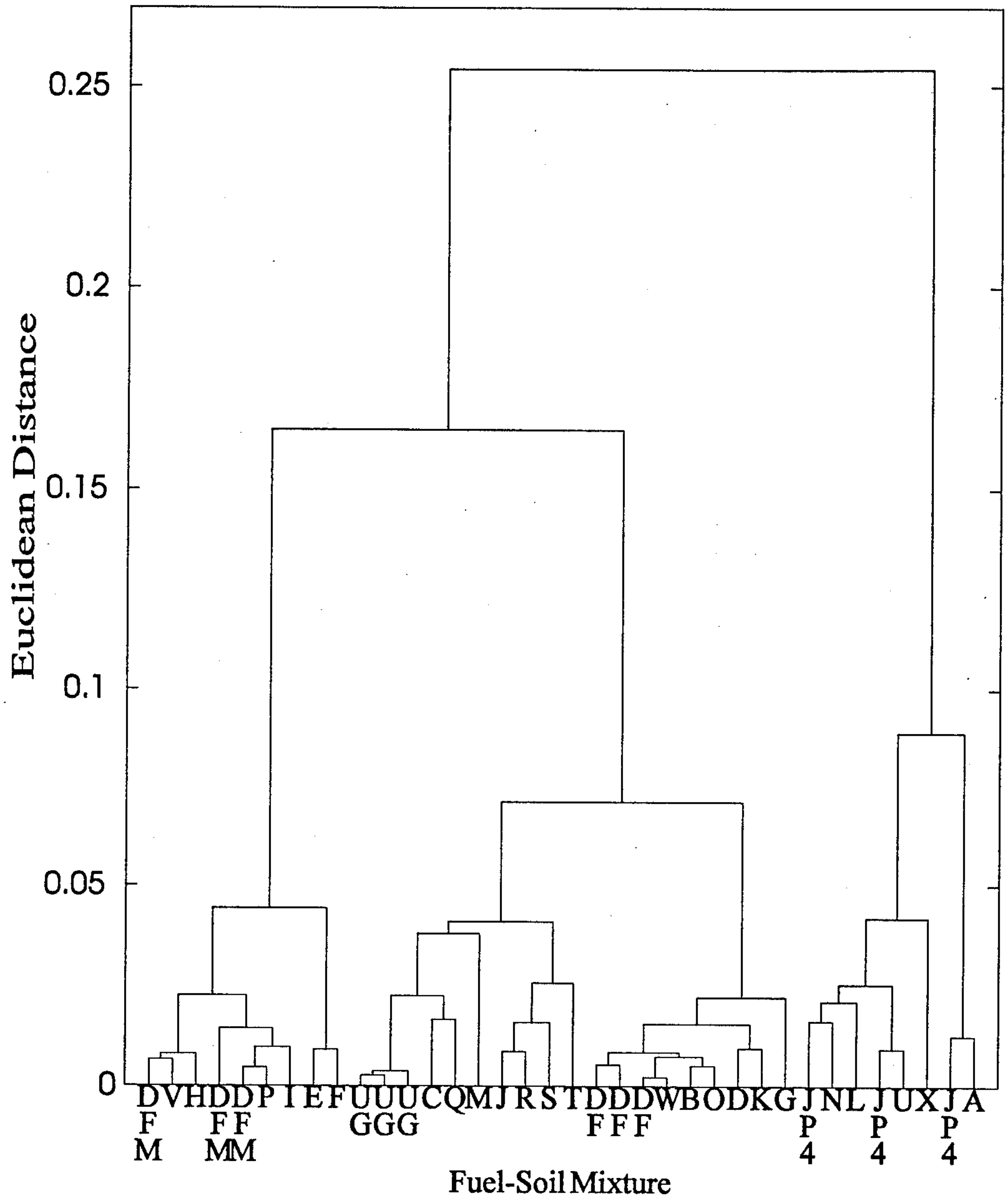


Figure 5.22. Average linkage dendrogram of fuels using three-mode PFA factors. Normalized mixture-mode factors. Sapphire window (interference) factor removed using rank five 3M-NNALS. Similarity measure is Euclidean distance between descriptor vectors. The known spectra are indicated by their fuel abbreviations and the unknowns by their letter designators.

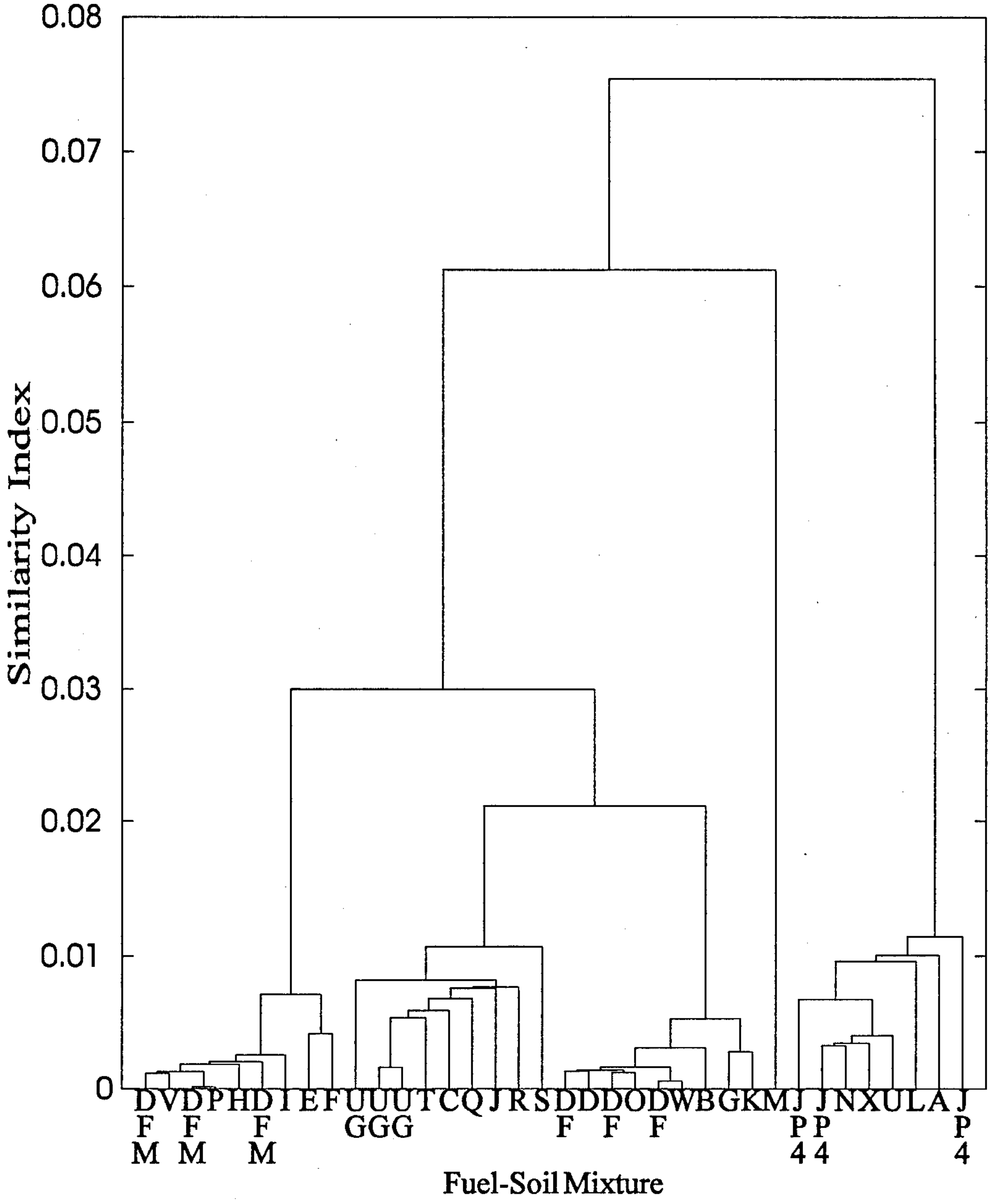


Figure 5.23. Single linkage dendrogram of time-integrated fuel spectra. The angle cosine of the spectral vectors subtracted from unity was used as the measure of similarity. Sapphire window (interference) factor, obtained from rank five 3M-NNALS decomposition, was removed before integration. The known spectra are indicated by their fuel abbreviations and the unknowns by their letter designators.

not surprising based on the results in the preceding section. One can easily see four groups using a similarity index of approximately 0.015 as a cutoff.

Based on the similarity index values in Figure 5.23, one might get the mistaken impression that the spectra are almost identical except for very subtle variations. This arises because the single linkage method finds the minimum distance or similarity between groups as they are established. In reality, the angle cosines (correlations) between different fuel spectra can be quite small, as low as 0.45 for DFM and JP4. Figure 5.23 does correctly indicate the general relative similarity between the fuel spectra, i.e., UG and DF have the most similar spectra, and DFM is more similar to UG and DF than JP4.

Figure 5.24 contains the complete linkage dendrogram for the same data depicted in Figure 5.23. Note that the groups have greater separation than in the single linkage model. Unknown M also does not appear to be as well separated from the other groups. These effects result from the choice of maximum distance or similarity between groups as opposed to the minimum in the single linkage method. The movement of Unknown M in relation to the other groups as the method is changed indicates of the spread of the angle cosines within the groups.

While the groups seem to have a large variance, given the large changes in the similarity indices from single linkage to complete linkage and the movement of Unknown M, the groups are well-defined and well-separated. This interpretation arises from the formation of the same groups regardless of method chosen. If different groups were formed as the method of HCA changed, which is common, then a very judicious choice of groups would be required.

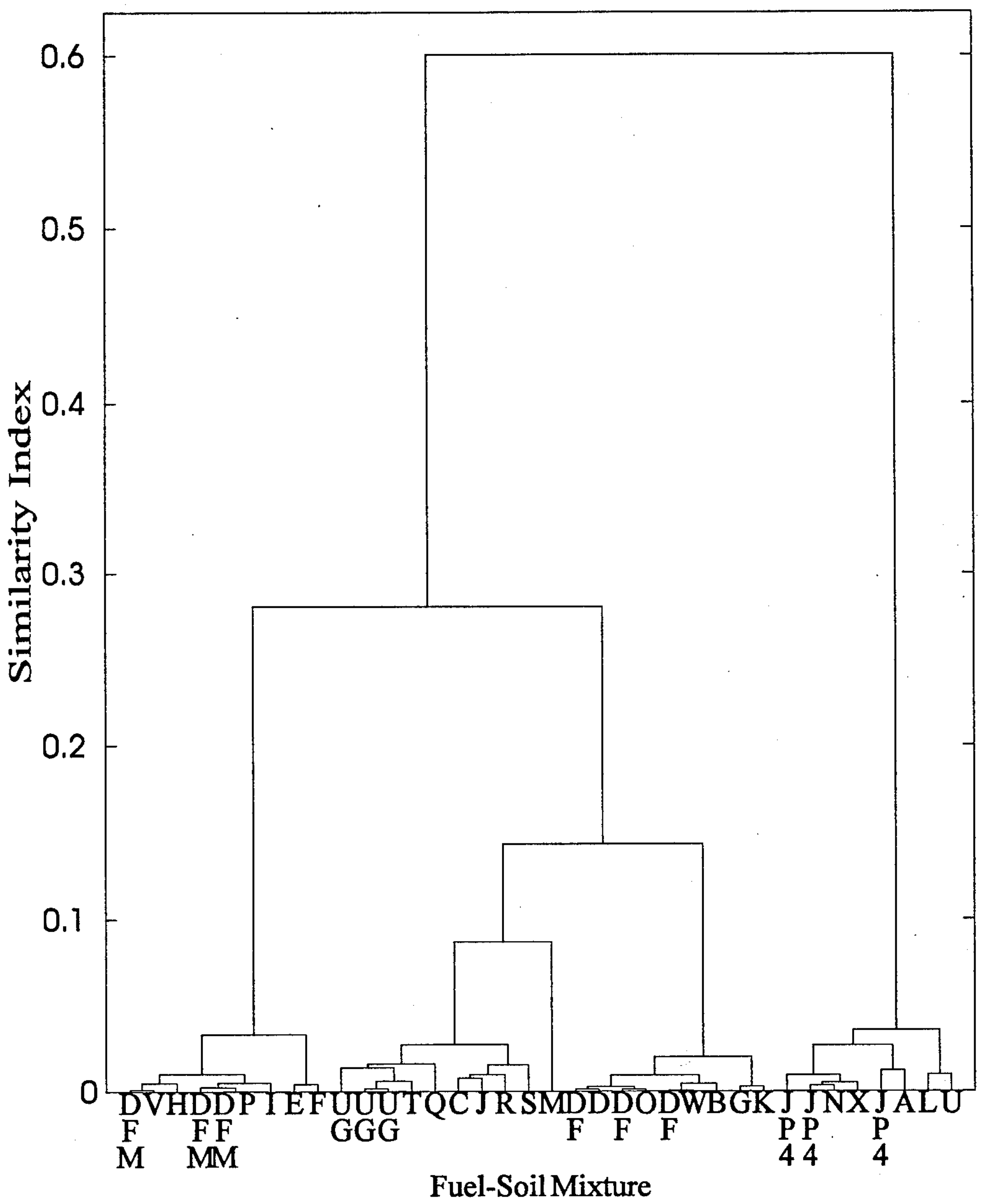


Figure 5.24. Complete linkage dendrogram of time-integrated fuel spectra. The angle cosine of the spectral vectors subtracted from unity was used as the measure of similarity. Sapphire window (interference) factor, obtained from rank five 3M-NNALS decomposition, was removed before integration. The known spectra are indicated by their fuel abbreviations and the unknowns by their letter designators.

Figure 5.25 contains the average linkage dendrogram for the data depicted in Figures 5.23 and 5.24. Once again, the groups are nicely separated and distinct. A similarity index cutoff of approximately 0.4 separates the four fuels, leaving Unknown M isolated.

Based on the results presented, HCA is an excellent method of grouping fuel mixtures. The average linkage method of clustering is probably as good a choice as any, although the single and complete linkage methods should be examined in each case. The angle cosine of the time-integrated fluorescence spectra is an excellent measure of similarity, and it is easy to apply to fluorescence emission data. This should be a valuable tool in site characterization.

Of course, as in classification, it is important to numerically remove any interferences before HCA. If the sapphire window component is not removed, many of the low concentration mixtures would not cluster correctly.

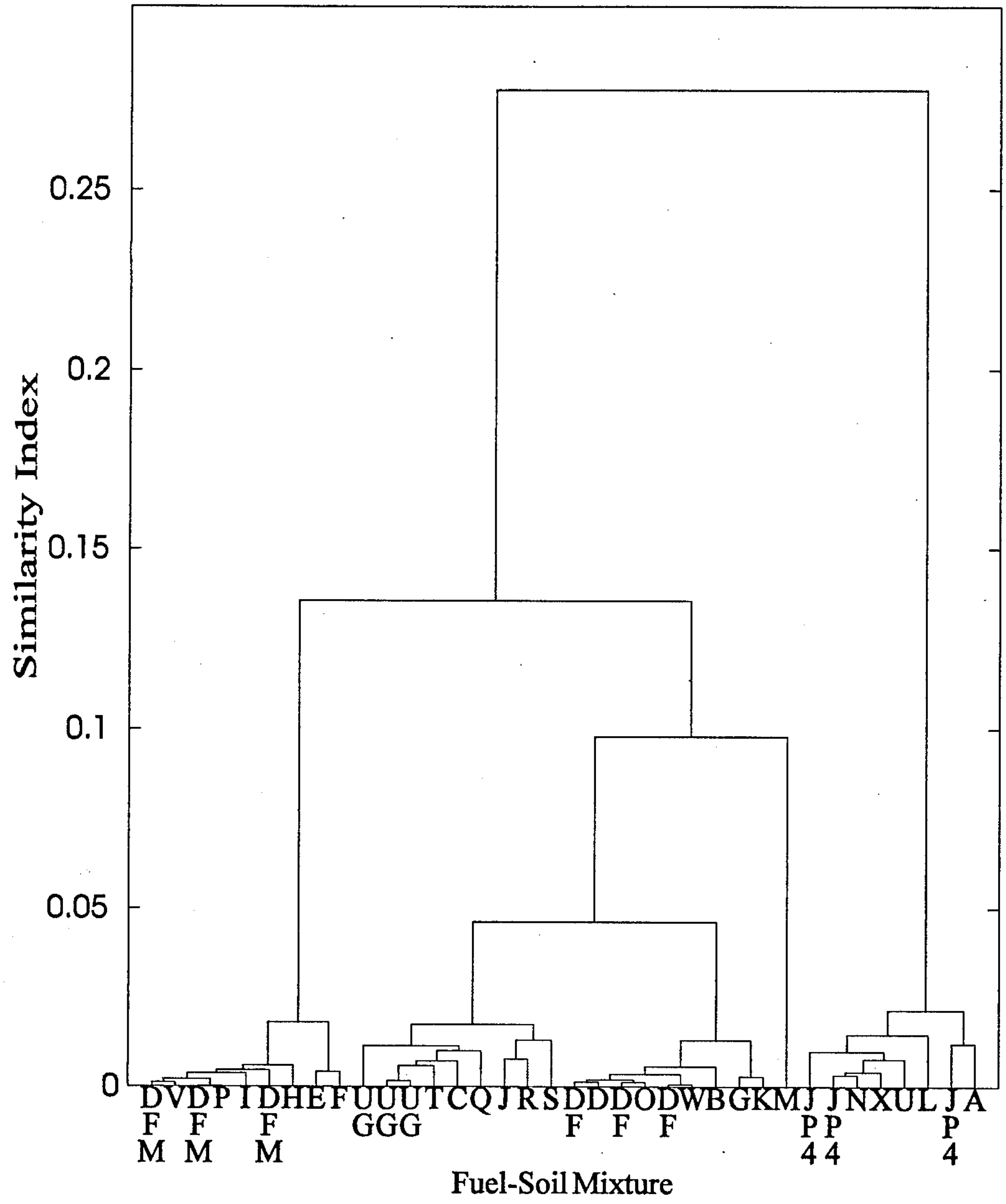


Figure 5.25. Average linkage dendrogram of time-integrated fuel spectra. The angle cosine of the spectral vectors subtracted from unity was used as the measure of similarity. Sapphire window (interference) factor, obtained from rank five 3M-NNALS decomposition, was removed before integration. The known spectra are indicated by their fuel abbreviations and the unknowns by their letter designators.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary of the Present Work

This research represents the first application of TLD and global analysis methods to a WTM-concentration 3-array and a TREEM. This is also the first use of TLD methods on highly complicated data in the form of fluorescence of fuels. Previous works have applied TLD methods to EEM-concentration/quencher level arrays[36,39,62] and EEFAs[59-61] of solutions with only a few components. TLD of WTM based data gives greater flexibility than EEM based data since it can be fit to a well-established model, exponential decay, which can be utilized in a three-mode global analysis scheme. WTM data typically have better S/N than EEFA data because optical switches used in phase-modulated instruments do not provide perfect modulation.

Global analysis of fluorescence decay time-mode data has been used to decompose WTMs of two component mixtures;[53] however, global analysis of three-mode data has not been performed.

The efficacy of classification and clustering algorithms applied to reduced forms of three-mode fuel fluorescence data was also demonstrated.

Analyses of Data Set One illustrate the capabilities of various rank estimation and profile extraction techniques with low rank data. Data Set One consists of seven WTMs. The WTMs were the fluorescence response of six solutions containing varying amounts of fluorene, pyrene, and naphthalene in water and a water blank. Including the water Raman scatter as a component, the 3-array has an *a priori* rank of the 3-array of four.

Excellent rank estimates of the matrix summed WTMs of Data Set One were obtained using Malinowski's factor indicator function, *F*-test, and first-lag autocorrelation coefficient of the wavelength-mode eigenvectors. The first-lag autocorrelation coefficient of the time-mode eigenvectors did not give good rank estimates. However, time-mode based estimates could be improved by using a cutoff value greater than +0.5, which was used for the wavelength mode. This is the first use of matrix summation of the WTMs before rank estimation reported. It represents a significant improvement of the capabilities of rank estimation of 3-arrays.

The four component WTM-concentration 3-array in Data Set One was easily decomposed with an EBP, 3M-ALS, and 3M-NNALS. 3M-ALS and 3M-NNALS offered a slight improvement to the EBP result. 3M-NNALS offered a minor computation time improvement over 3M-ALS.

Profile extraction using global analysis on Data Set One provided good results when seeking three nonzero lifetimes and including a zero lifetime scatter component. However, global analysis did not perform well when attempting to also extract the profile of a very weak interference, which was discovered in the three-mode analyses.

Data Set Two is a four-component TREEM of a single solution of fluorene, naphthalene, carbazole, and phenanthrene. The spectra of carbazole and phenanthrene are heavily overlapped, and the lifetimes of naphthalene and phenanthrene are very similar. These factors made for a difficult decomposition problem.

Performance of rank estimation methods were essentially the same as for Data Set One. A matrix representing the compressed 3-array was produced by matrix summation of the four WTMs obtained at the excitation wavelengths utilized to generate the TREEM. Excellent rank estimates were obtained using Malinowski's factor indicator function, *F*-test, and first-lag autocorrelation coefficient of the wavelength-mode eigenvectors. The first-lag autocorrelation coefficient of the time-mode eigenvectors did not give good rank estimates using +0.5 as a cutoff value.

TLD was performed on this data using an EBP, 3M-ALS, and 3M-NNALS. The EBP and 3M-ALS were unable to provide realistic factor profiles for this data. This was probably the result of the similarity in lifetimes of naphthalene and phenanthrene. 3M-NNALS produced very good results, generating recognizable factors in one-tenth of the time required by 3M-ALS. 3M-NNALS provides substantially improved results over 3M-ALS, a notable advancement of three-mode decomposition.

Global analysis, performed using the methodology employed with Data Set One, failed to produce a satisfactory result for any number of factors. The algorithm either failed to obtain recognizable factors or terminated with two nearly identical lifetimes. The latter condition resulted in an ill-conditioned solution of the excitation-emission mode.

A hybrid form of analysis, which combines global analysis with NNALS, provided very promising results that were nearly identical to 3M-NNALS. However, the algorithm was very time intensive. This utilization of the time-mode information in conjunction with three-mode analysis is unique. We believe this methodology has great potential as an analysis technique because there is a model in the time mode which does not exist in the excitation or emission modes.

The final set of data examined in this research was composed of WTMs of fuels on soil matrices. This data contained an interference in the form of sapphire window fluorescence. The goal of analysis was to use TLD to remove the interference, and, in the process, generate descriptors of the data that could be used in a classification scheme.

Rank estimation was performed for this data in the same manner as in the two previous data sets. The rank estimates were smaller than expected, considering the number of fluorescing species in fuels.

TLD using an EBP was not successful, yielding complex eigenvectors or chemically meaningless factors. 3M-ALS performed better than the EBP, but it produced multiple factor degeneracies and took a great deal of computation time. 3M-NNALS performed best of the three TLD procedures. It generated factors which may be interpreted as real factors, although no assignment was attempted. Computation time was at least two orders of magnitude faster than 3M-ALS. The rank five decomposition provided the best match for the sapphire window factor.

Classifications of the unknown fuels using the mixture-mode factors from 3M-NNALS were very good. Using all of the rank five factors as descriptors, 22 out of 24 unknown fuels were correctly classified. Omitting the suspected window factor and using the remaining four factors as descriptors, 22 out of 24 unknowns were classified correctly. One of the unknowns was misclassified in both cases. In all cases of misclassification, the fuels were low concentration and were usually on the soil matrix with the poorest fluorescence response.

Classification rates were improved, often to 100%, when the window factor was subtracted from the 3-array. After performing eigenanalysis on the 3-array or the time-integrated spectra of the fuels, the mixture-mode eigenvectors were used as descriptors of the fuels.

HCA was performed utilizing metrics computed from the various descriptors described in the preceding paragraph and one minus the angle cosine as measures of similarity. HCA was a good method of grouping the fuel mixtures. Using the angle cosine as the measure of similarity is intuitive and gave excellent cluster results. Forming linkages with the all linkage methods worked well, and each should be utilized for each application of HCA.

6.2 Recommendations for Future Work

Future efforts in this area should be directed at evaluating more fuels and fuel types. These analysis techniques may be able to detect subtle differences in fuels, but this must be explored.

To understand the complexities of fuel component interactions, mixtures of simulated fuels can be prepared and analyzed using these fluorescence techniques. Simulated fuels could easily be prepared in a progressive fashion, allowing examination of the interactions as they develop. Using this methodology, phenomena such as the low rank of the fuel WTMs may be addressed.

Oxygen quenching can be problematic in time-mode fluorescence spectroscopy. In the presence of oxygen, the intensity of the fluorescence is diminished and the lifetime shortened. If two samples of the same fuel undergo different levels of quenching, their components will have different lifetimes, but identical emission modes. This will make for a difficult analysis problem because the time-mode will have a higher rank than the emission mode. Restricted Tucker models[37] are designed to deal with problems of this type. Current programs can be easily modified and tested to perform this type of analysis.

The 3M-NNALS routine works by setting all negative elements in a matrix to zero after each iteration. This rather abrupt method may lead to trapping the solution in a local minimum. A more amenable procedure would involve setting the negative elements progressively less negative each iteration. For example, reduce the negative elements by 1% on the first iteration; and on successive iterations, reduce them by an additional 1%. In this way, the unrealistic solutions are guided gently back into the region of real solutions.

Since fluorescence is easily capable of higher orders than used in the present work, consideration should be given to acquisition and analysis of fourth order data. Excitation-emission-time-concentration data could be generated at the present time with no modifications to existing equipment. Advantages of this type of data over third-order data are not known. However, an additional mode of data may allow degeneracies in one mode to be lifted in another, for example, quenching.

Time-mode data is unique because an entire vector of data can be expressed as a single variable, lifetime, for a single component. Greater effort should be directed at developing the three-mode global analysis. One method which may be of interest would force the time-mode to fit the convolutional model of equation 2.11 after each iteration of a 3M-ALS or 3M-NNALS sequence. This may generate more realistic solutions in some cases.

These physical methods and fluorescence analysis techniques should be applied to areas outside of environmental analysis. Medical diagnostics is one area that has been exploring fluorescence. Advances in environmental analysis may be readily applied to diagnosis of cancer and cardiovascular disease and medical monitoring.

Classification of healthy and diseased tissue utilizing third- or higher-order data with one mode of data as a healthy/diseased-mode (analogous to the fuel type mode) is easily conceivable. This type of analysis would facilitate identification of the biochemical indicators of disease.

REFERENCES

1. Elving, P. J.; Kienitz, H. *Treatise on Analytical Chemistry, Part I: Theory and Practice*, 2nd ed.; Kolthoff, I. M.; Elving, P. J., John Wiley and Sons, Inc.: New York, 1978; Vol. 1, Chapter 3.

2. Skoog, D. A.; West, D. M.; Holler, James F. *Fundamentals of Analytical Chemistry*; New York: Saunders College Publishing, 1988.

3. Bicking, C. A. *Treatise on Analytical Chemistry, Part I: Theory and Practice*, 2nd ed.; Kolthoff, I. M.; Elving, P. J., John Wiley and Sons, Inc.: New York, 1978; Vol. 1, Chapter 6.

4. Putzig, C. L.; Leurgers, M. A.; McKelvey, M. L.; Mitchell, G. E.; Nyquist, R. A.; Papenfuss, R. R.; Yurga, L. *Anal. Chem.* **1994**, *66*(12), 26R-66R.

5. Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*(15), A782-A791.

6. Thomas, E. V. *Anal. Chem.* **1994**, *66*(15), 795A-804A.

7. Kowalski, B. R.; Seasholtz, M. B. *J. Chemom.* **1991**, *5*, 129-145.

8. Warner, I. M.; Patonay, G.; Thomas, M. P. *Anal. Chem.* **1985**, *57*(3), 463A-483A.

9. Skoog, D. A.; Leary, J. J. *Principles of Instrumental Analysis*; Saunders College Publishing: Fort Worth, 1992.

10. Mitchell, B. C. Ph.D. dissertation, Duke University, Durham, NC, 1993.

11. Kruskal, J. B. *Linear Algebra Appl.* **1977**, *18*, 95-138.

12. Kruskal, J. B. *Multiway Data Analysis*; Coppi, R.; Bolasco, S., Editors; North-Holland: Amsterdam, 1989; pp 7-18.

13. Gillispie, G. D.; St. Germain, R. W.; Klingfus, J. L. *Proceedings of the 1993 U.S. EPA/A&WMA International Symposium: Field Screening Methods for Hazardous Wastes and Toxic Chemicals;* Air and Waste Management Association: Pittsburgh, PA, 1993; pp 793-805.

14. Top 20 Hazardous Substances: ATSDR/EPA Priority List (1993) ATSDR: http://atsdr1.atsdr.cdc.gov:8080/cxcx3.html, 1993.

15. Diehl, J. W.; Finkbeiner, J. W.; DiSanzo, F. P. *Anal. Chem.* **1993**, *65*, 2493-2496.

16. Cline, P. V.; Delfino, J. J.; Rao, P. S. C. *Environ. Sci. Technol.* **1991**, *25*(5), 914-920.

17. Lee, L. S.; Hagwall, M.; Delfino, J. J.; Rao, P. S. C. *Environ. Sci. Technol.* **1992**, *26*(11), 2104-2110.

18. Polycyclic Aromatic Hydrocarbons ATSDR: http://atsdr1.atsdr.cdc.gov:8080/ToxProfiles/PHS/Polycyclic_aromatic_hydrocarbons_[PAHs].1990.html, 1990.

19. Keith, L. H. *Environmental Sampling and Analysis: A Practical Guide*; Lewis Publishers: Chelsea, MI, 1991.

20. *Test Methods for Evaluating Solid Waste: Physical/Chemical Methods*; U.S. EPA, Office of Solid Waste and Emergency Response: Washington, DC, 1986.

21. Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier: Amsterdam, 1988.

22. Brereton, R. A. *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*; Ellis Horwood: New York, 1990.

23. Parker, C. A. *Photoluminescence of Solutions*; Elsevier: Amsterdam, 1968.

24. McGlynn, S. P.; Azumi, T.; Kinoshita, M. *Molecular Spectroscopy of the Triplet State*; Prentice-Hall: Englewood Cliffs, NJ, 1969.

25. Demas, J. N. *Excited State Lifetime Measurements*; Academic Press: New York, 1983.

26. Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*; Plenum Press: New York, 1983.

27. Farden, D. C. *Fundamentals of Signals and Linear Systems*; Fargo: NDSU Press, E.E. Dept., 1992.

28. Gillispie, G. D. *Structure-Property Relations in Polymers : Spectroscopy and Performance*; Urban, M.; Craver, C. D., Eds.; Advances in Chemistry Series 236; American Chemical Society: Washington, DC, 1993; Vol. 236, pp 89-127.

29. Schneckenburger, H.; Sidlitz, H. K.; Eberz, J. *J. Photochem. Photobiol., B:Biol.* **1988**, *2*, 1-19.

30. Meidinger, R. F. M.S. thesis, North Dakota State University, Fargo, 1993.

31. *TDS 620 & 640 Digitizing Oscilloscope User Manual;* Tektronix, Inc.: Beaverton, OR, 1992.

32. McGown, L. B.; Bright, F. V. *Anal. Chem.* **1984**, *56*(13), 1400A-1416A.

33. Malinowski, E. R. *Factor Analysis in Chemistry*; John Wiley and Sons, Inc.: New York, 1991.

34. Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 247-263.

35. Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, 2, 265-280.

36. Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1990**, *4*, 29-45.

37. Smilde, A. K.; Wang, Y. D.; Kowalski, B. R. *J. Chemom.* **1994**, *8*(1), 21-36.

38. Henshaw, J. M.; Burgess, L. W.; Booksh, K. S.; Kowalski, B. R. *Anal. Chem.* **1994**, *66*(20), 3328-3336.

39. Leurgans, S. E.; Ross, R. T. *Statistical Science* **1992**, *7*(3), 289-319.

40. Andersson-Engels, S.; Johansson, J.; Svanberg, K.; Svanberg, S. *Photochem. Photobiol.* **1991**, *53*(6), 807-814.

41. Geladi, P.; Kowalski, B. R. *Analytica Chimica Acta* **1986**, *185*, 1-17.

42. Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193-1202.

43. Wilson, B. E.; Kowalski, B. R. *Anal. Chem.* **1989**, *61*(20), 2277-2284.

44. Ohman, J.; Geladi, P.; Wold, S. *J. Chemom.* **1990**, *4*(2), 79-90.

45. Ohman, J.; Geladi, P.; Wold, S. *J. Chemom.* **1990**, *4*(2), 135-146.

46. Malinowski, E. R. *J. Chemom.* **1988**, *3*, 49-60.

47. Shrager, R. I.; Hendler, R. W. *Anal. Chem.* **1982**, *54*(7), 1147-1152.

48. Spjotvoll, E.; Martens, H.; Volden, R. *Technometrics* **1982**, *24*(3), 173-180.

49. Tauler, R.; Casassas, E.; Izquierdo-Ridorsa, A. *Anal. Chim. Acta* **1991**, *248*, 447-458.

50. Tauler, R.; Kowalski, B.; Fleming, S. *Anal. Chem.* **1993**, *65*(15), 2040-2047.

51. Tauler, R.; Izquierdo-Ridorsa, A.; Casassas, E. *Chemom. Intell. Lab. Sys.* **1993**, *18*, 293-300.

52. Tauler, R.; Barcelo, D. *Trends Anal. Chem.* **1993**, *12*(8), 319-327.

53. Knorr, F. J.; Harris, J. M. *Anal. Chem.* **1981**, *53*(2), 272-276.

54. Beechem, J. P. *Numerical Computer Methods*; Brand, L.; Johnson, M. L., Editors; Methods in Enzymology 201; Academic Press: San Diego, 1992; pp 37-54.

55. Henry, E. R.; Hofrichter, J. *Numerical Computer Methods*; Brand, L.; Johnson, M. L., Editors; Methods in Enzymology 201; Academic Press: San Diego, 1992; pp 129-192.

56. Bevington, P. R.; Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences*; McGraw-Hill, Inc.: New York, 1992.

57. Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*(3), 617-633.

58. Basilevsky, A. *Statistical Factor Analysis and Related Methods: Theory and Applications*; John Wiley and Sons, Inc.: New York, 1994.

59. Millican, D. W.; McGown, L. B. *Anal. Chem.* **1989**, *61*, 580-583.

60. Millican, D. W.; McGown, L. B. *Anal. Chem.* **1990**, *62*, 2242-2247.

61. Burdick, D. S.; Tu, X. M.; McGown, L. B.; Millican, D. W. *J. Chemom.* **1990**, *4*, 15-28.

62. Leurgans, S. E.; Ross, R. T.; Abel, R. B. *SIAM J. Matrix Anal. Appl.* **1993**, *14*(4), 1064-1083.

63. Mitchell, B. C.; Burdick, D. S. *J. Chemom.* **1994**, *8*, 155-168.

64. Kruskal, J. B.; Harshman, R. A.; Lundy, M. E. *Multiway Data Analysis*; Coppi, R.; Bolasco, S., Editors; North-Holland: Amsterdam, 1989; pp 115-122.

65. Booksh, K., personal communication, 1994.

66. Hecht, H. G. *Mathematics in Chemistry*; Prentice Hall: Englewood Cliffs, NJ, 1990.

67. James, M. *Classification Algorithms*; John Wiley and Sons, Inc.: New York, 1985.

68. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press, Inc.: Boston, 1990.

69. Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 1992.

70. Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis* ; John Wiley and Sons, Inc.: New York, 1983.

71. Killeen, T. J.; Eastwood, D.; Hendrick, M. S. *Talanta* **1981**, *28*, 1-6.

72. Li, S.; Hamilton, J. C.; Gemperline, P. J. *Anal. Chem.* **1992**, *64*, 599-607.

73. Berlman, I. B. *Handbook of Fluorescence Spectra of Aromatic Molecules*; Academic Press: New York, 1971.

74. Ware, W. R. *Photochemistry in Organized and Constrained Media*; Ramamurthy, V., Editor; VCH Publishers, Inc.: New York, 1991; pp 563-602.

75. Hentschel, C. *Fiber Optics Handbook;* 3rd ed. Hewlett-Packard Co.: Boblingen, Germany, 1989.

76. Melles Griot *Optics Guide*; 5; Melles Griot: Irvine, CA, 1990.

77. Mitchell, B. C.; Burdick, D. S., personal communication, 1993.

78. Eastwood, D., personal communication, 1993.

79. Rummel, R. J. *Applied Factor Analysis*; Northwestern University Press: Evanston, IL, 1970.

80. Harman, H. H. *ModernFactor Analysis*; University of Chicago Press: Chicago, 1976.

81. Tallarida, R. J. *Pocket Book of Integrals and Mathematical Formulas*; CRC Press: Boca Raton, 1991.

82. Tucker, L. R. *Psychometrika* **1966**, *31*(3), 279-311.

83. Kroonenberg, P. M.; de Leeuw, J. *Psychometrika* **1980**, *45*(1), 69-97.

APPENDIX A. DEFINITIONS AND NOTATION

Throughout this thesis, I used the conventions drawn from the chemometric literature for displaying vectors and matrices.[33,58,79-81] Scalar quantities are represented by lower case letters. Scalars which are elements of a vector or matrix are labeled by subscript indices indicating row and column. Column vectors are indicated by bold lower case letters, e.g., **a**. Row vectors are denoted by a prime, e.g., $\mathbf{a}'$. Matrices are represented by bold upper case letters, e.g., **A**. Therefore,

a_{ij} is a scalar (the subscript denotes the $\{ij\}$th element of the matrix **A**).

$$\mathbf{a}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{\mathrm{m}j} \end{bmatrix}$$ is the *j*th column vector (in m-coordinate space).

$\mathbf{a}_i' = [a_{i1}\ a_{i2}\ \cdots\ a_{in}]$ is the ith row vector (in n-coordinate space).

Definition A.1. (Inner Product) The inner product (also called the *dot product* or *scalar product*) of two vectors is

$$\mathbf{a}\cdot\mathbf{b} = \mathbf{a}'\mathbf{b} = [a_1\ a_2\ \cdots\ a_{\mathrm{n}}]\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{\mathrm{n}} \end{bmatrix} = a_1b_1 + a_2b_2 + \cdots + a_{\mathrm{n}}b_{\mathrm{n}} \qquad \text{(A.1)}$$

Definition A.1. Note 1. The inner product is a scalar, and it requires that both vectors have the same number of elements.

Definition A.1. Note 2. The properties of the inner product are preserved whether the vectors are row or column vectors, i.e.,

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}' \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{b}'$$

Definition A.1. Note 3. (Vector Norm) The length (or norm) of a vector is given by the square root of the inner product of the vector with itself:

$$\|\mathrm{a}\| = \left(\sum a_i^2\right)^{\frac{1}{2}} = \sqrt{\mathrm{a}'\mathrm{a}} = \sqrt{\mathrm{a} \cdot \mathrm{a}} \qquad \text{(A.2)}$$

Definition A.2. (Normal Vector) A normal vector has unit length. That is, for the normal vector **a**,

$$\|\mathrm{a}\| = 1.$$

Definition A.2. Note 1. A vector may be *normalized* by setting its length equal to one. This is done by dividing the vector by its length, e.g.,

$$\frac{\mathbf{a}}{\|\mathrm{a}\|}.$$

Definition A.3. (Orthogonal Vectors) Two vectors are orthogonal if their inner product is equal to zero (0).

Definition A.4. (Orthonormal Vectors) Two vectors are orthonormal if their inner product is equal to zero and their lengths are equal to one (1).

Definition A.5. (Matrix Dimension) A matrix that has *m* rows with *n* elements in each row is said to have matrix dimensions *m* by *n*. A convenient shorthand for representing the

m by n matrix $\mathbf{A}$ is $\mathbf{A}_{m\times n}$, so

$$\mathbf{A} = \mathbf{A}_{3\times 2} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \text{ is 3 by 2 a matrix.}$$

Definition A.5. Note 1. In many texts, this property of a matrix is called the *order* of the matrix. However, the term *order* is used to describe a variety of properties of matrices and arrays in various texts and publications and here.

Definition A.6. (Matrix Addition) Matrix addition is the addition of the corresponding elements of matrices of identical dimensions. For example, the sum of matrices $\mathbf{A}_{3\times 2}$ and $\mathbf{B}_{3\times 2}$ is

$$\mathbf{A}_{3\times 2} + \mathbf{B}_{3\times 2} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \\ a_{31}+b_{31} & a_{32}+b_{32} \end{bmatrix}$$

Definition A.6. Note 1. Matrix addition of matrices of different dimensions is undefined.

Definition A.6. Note 2. Vector addition is completely analogous to matrix addition where one of the matrix dimensions is one (1).

Definition A.7. (Scalar Multiplication) Multiplication of a scalar by a matrix is the product of each element of the matrix by the scalar multiplier. The product of a and $\mathbf{B}_{3\times 2}$ is

$$a\mathbf{B}_{3\times 2} = a\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} ab_{11} & ab_{12} \\ ab_{21} & ab_{22} \\ ab_{31} & ab_{32} \end{bmatrix}$$

Definition A.7. Note 1. Scalar multiplication of a vector is completely analogous to scalar multiplication of a matrix.

Definition A.8. (Matrix Multiplication) Multiplication of two matrices, denoted **AB**, is accomplished by computing the inner product of the *i*th row of the left matrix (**A**)with the *j*th column of the right matrix(**B**). Matrix multiplication requires that the two matrices be *conformable*, that is, the number of columns of the left matrix must equal the number of rows of the right matrix. Therefore, if $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ are to be multiplied, then *n* must equal *p*. If $n \neq p$, then multiplication is not defined. For example,

$$\mathbf{AB}=\begin{bmatrix}2 & 0 & 1\\ 1 & 4 & 3\end{bmatrix}\begin{bmatrix}2 & 4 & 3\\ 2 & 1 & 1\\ 2 & 3 & 1\end{bmatrix}=\begin{bmatrix}4+0+2 & 8+0+3 & 6+0+1\\ 2+8+6 & 4+4+9 & 3+4+3\end{bmatrix}=\begin{bmatrix}6 & 11 & 7\\ 16 & 17 & 10\end{bmatrix}$$

Definition A.9. (Matrix Rank) The rank of a matrix is the maximum number of linearly independent column (or row) vectors in the matrix.

Definition A.9. Note 1. The rank of a matrix is the true underlying order of that matrix. Rank is also defined as the order of the largest matrix with a nonzero determinant formed by deleting rows and columns of the original matrix. The maximum rank of the matrix $\mathbf{A}_{m\times n}$ is the minimum value of *m* and *n*. The rank of $\mathbf{A}_{m\times n}$ is denoted by $R\ (\mathbf{A}_{m\times n})$.

Definition A.9. Note 2. A matrix whose rank is less than minimum of its dimensions is said to be rank deficient.

Definition A.9. Note 3. A zero matrix (a matrix all of whose elements are zero) has rank 0.

Definition A.10. (Square Matrix) A square matrix has the same number of rows and columns. Therefore, the matrix $\mathbf{A}_{m \times n}$ is square if $m=n$.

Definition A.11. (Matrix Transpose) The transpose of a matrix is obtained by interchanging the rows and columns of the matrix. Transpose is indicated by a prime, so,

$$\text{If } \mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 5 & 4 & 7 \end{bmatrix} \text{ then } \mathbf{A}' = \begin{bmatrix} 2 & 5 \\ 1 & 4 \\ 3 & 7 \end{bmatrix}$$

Definition A.12. (Diagonal Matrix) A diagonal matrix is a square matrix with zero elements except along the principal diagonal. Any matrix, **A**, is diagonal if $a_{ij} = 0$ for all $i \neq j$, and if at least one $a_{ij} \neq 0$ when $i = j$. Therefore,

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix} \text{ is a diagonal matrix.}$$

Definition A.13. (Identity Matrix) An identity matrix, **I**, is a diagonal matrix with ones along the principal diagonal. Since the identity matrix is always a square matrix, its dimensions are indicated by a single subscript, $\mathbf{I}_n$. The matrix product of any matrix with the identity matrix is itself.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\mathbf{A}_{3\times2}\mathbf{I}_2 = \mathbf{A}_{3\times2}$$

Definition A.14. (Matrix Inverse) The inverse of an *n* by *n* (i.e., a square matrix) is another matrix such that the product of the two matrices is an identity matrix. The inverse of a matrix is denoted by a superscript (-1). An invertible matrix is said to be *nonsingular.*

$$\mathbf{A}_{3\times3}^{-1}\mathbf{A}_{3\times3} = \mathbf{I}_3$$

Definition A.14. Note 1. A rank deficient matrix will not be invertible since it will have a determinant equal to zero. Rank deficient matrices are also called *singular* matrices.

Definition A.15. (Matrix Pseudoinverse) The pseudoinverse of a matrix is the least squares solution to a set of linear equations. It is designated by a superscripted cross (+).

$$\begin{aligned}
&\mathbf{b} = \mathbf{Ax} \\
&\mathbf{A'b} = \mathbf{A'Ax} \\
&(\mathbf{A'A})^{-1}\mathbf{A'b} = (\mathbf{A'A})^{-1}\mathbf{A'Ax} \\
&(\mathbf{A'A})^{-1}\mathbf{A'b} = \mathbf{Ix} = \mathbf{x} \\
&\mathbf{A}^{+}\mathbf{b} = \mathbf{x} \\
&\mathbf{A}^{+} \equiv (\mathbf{A'A})^{-1}\mathbf{A'} \equiv \mathit{pseudoinverse}(\mathrm{A})
\end{aligned}$$

Definition A.15. Note 1. The pseudoinverse is usually used when **A** is a rectangular matrix (i.e., *m* by *n*), since **A** has no true inverse.. Also, the condition $m > n$ must be satisfied or the matrix $(\mathbf{A'A})$ will be singular.

Definition A.16. (Orthonormal Matrix) A matrix whose column vectors are mutually orthonormal is a column-wise orthonormal matrix. It has the convenient property:

$$\mathbf{A'A} = \mathbf{I}.$$

Likewise, a matrix whose row vectors are mutually orthonormal is a row-wise orthonormal matrix, and

$$\mathbf{AA'} = \mathbf{I}.$$

Definition A.16. Note 1. (Orthogonal Matrix) A matrix whose column vectors are mutually orthogonal is a column-wise orthogonal matrix. It has the property:

$$\mathbf{A}'\mathbf{A}=\mathbf{D}$$

where $\mathbf{D}$ is a diagonal matrix.

Likewise, a matrix whose row vectors are mutually orthogonal is a row-wise orthogonal matrix, and

$$\mathbf{A}\mathbf{A}'=\mathbf{D}.$$

Definition A.17. (Matrix Square Root) Let **Y** be the matrix square root of **X**, then,

$$\mathbf{Y}\mathbf{Y}=\mathbf{X}.$$

and,

$$\mathbf{Y}=\mathbf{X}^{1/2}.$$

Definition A.18. (Matrix Norm) The norm of a matrix is a scalar that gives a measure of the magnitude of the elements of the matrix. Here the matrix norm is defined as:

$$\|\mathrm{A}\| = \sqrt{\left(\sum\sum a_{ij}^2\right)} \tag{A.3}$$

Definition A.19. (Angle Cosine) The cosine of the angle, ϕ, between two vectors **a** and **b** (also called the angle cosine) is

$$\cos\phi = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\|\ \|\mathbf{b}\|} \tag{A.4}$$

Definition A.20. (*N*-way Array) An array is an ordered collection of scalars. An *N*-way array, or *N*-array, is an ordered collection of scalars which have indices that extend into *N*-dimensions. A matrix is a 2-array and is a rectangular array of scalar elements. Likewise, a 3-array is a rectangular parallelepiped array (a "*box*" or "*cube*") of scalar elements. Higher order arrays ("*hyperboxes*" or "*hypercubes*") extend into higher dimensional space and do not have a geometrical analog, but they are readily conceivable and do exist.

Definition A.20. Note 1. Many authors substitute the term "*mode*" for "*way*" when discussing N-arrays. Tucker[82] introduced this term to describe "a set of indices by which data might be classified." The two terms are synonymous.

Definition A.21. (Array Order) The order of an N-way array is the number of dimension into which the array extends, i.e., N. A scalar is defined as a zero-order array.

Using Definitions A.13 and A.14 then:

scalar $\equiv$ 0-array = array of order 0,

vector = 1-array = array of order 1,

matrix = 2-array = array of order 2,

box = 3-array = array of order 3,

hyperbox = (N>3)-array.

Definition A.22. (Array Unfolding) An N-array (N>2) can be rearranged so that it is in the form of a matrix or a partitioned matrix. When the different "*layers*" or "*slices*" or "*slabs*" of the N-array are placed into a single "*plane,*" this is referred to as the "*unfolded*" or "*flattened*" form of the array.

Definition A.22. Note 1. The unfolding operation can be performed in many ways and must be in the analyses discussed in this thesis.

Definition A.23. (Outer Product) The outer product (also regularly called the *direct product* or *tensor product*) of N vectors, **a**, **b**, **c**,... whose elements are a_i, b_j, c_k,..., generates an N-array whose elements are $a_i b_j c_k \ldots$. Outer product of vectors is indicated by $\otimes$, e.g., $\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} \otimes \ldots$

Definition A.24. (Outer Product Array) An outer product array is any array which can be expressed as an outer product.

Definition A.24. Note 1. The outer product of two vectors, **a** and **b,** is the outer product 2-array, **M**:

$$\mathbf{M} = \mathbf{a} \otimes \mathbf{b} = \mathbf{ab}' = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \end{bmatrix} = \begin{bmatrix} a_1b_1 & a_1b_2 & a_1b_3 & a_1b_4 \\ a_2b_1 & a_2b_2 & a_2b_3 & a_2b_4 \\ a_3b_1 & a_3b_2 & a_3b_3 & a_3b_4 \end{bmatrix} \quad \text{(A.5)}$$

Clearly, **M** is a rank 1 matrix since all of the columns (and rows) are linear combinations of the other columns (rows). A rank 1 matrix, i.e., outer product 2-array, is referred to as a *dyad.*

Definition A.25. The outer product of three vectors, $\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$, generates an outer product 3-array which is referred to as a *triad.* The general case of an outer product N-array produces an N-ad. The outer product of three vectors and the 3-array generated is displayed in Figure A.1 and represented algebraically in Equation A6. Unfolding any N-array is a straightforward extension of the three-way model shown in Figure A.1.

$$\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} \equiv$$

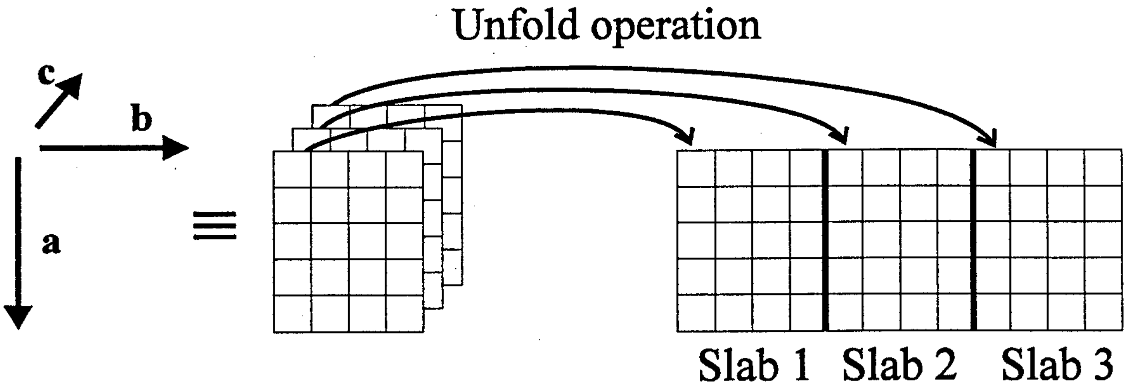


Figure A.1. Unfolding of outer product 3-array. The outer product of three vectors **a**, **b**, and **c** with dimensions 5, 4, and 3, respectively. The result is a 3-array with 3 (= c dimension), 5 by 4 slabs. Unfolding yields the 5 by 12 matrix.

Figure A.1, written in matrix form, is

$$\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} \otimes \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \otimes \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} =$$

$$\left[\begin{array}{cccc|cccc|cccc} a_1b_1c_1 & a_1b_2c_1 & a_1b_3c_1 & a_1b_4c_1 & a_1b_1c_2 & a_1b_2c_2 & a_1b_3c_2 & a_1b_4c_2 & a_1b_1c_3 & a_1b_2c_3 & a_1b_3c_3 & a_1b_4c_3 \\ a_2b_1c_1 & a_2b_2c_1 & a_2b_3c_1 & a_2b_4c_1 & a_2b_1c_2 & a_2b_2c_2 & a_2b_3c_2 & a_2b_4c_2 & a_2b_1c_3 & a_2b_2c_3 & a_2b_3c_3 & a_2b_4c_3 \\ a_3b_1c_1 & a_3b_2c_1 & a_3b_3c_1 & a_3b_4c_1 & a_3b_1c_2 & a_3b_2c_2 & a_3b_3c_2 & a_3b_4c_2 & a_3b_1c_3 & a_3b_2c_3 & a_3b_3c_3 & a_3b_4c_3 \\ a_4b_1c_1 & a_4b_2c_1 & a_4b_3c_1 & a_4b_4c_1 & a_4b_1c_2 & a_4b_2c_2 & a_4b_3c_2 & a_4b_4c_2 & a_4b_1c_3 & a_4b_2c_3 & a_4b_3c_3 & a_4b_4c_3 \\ a_5b_1c_1 & a_5b_2c_1 & a_5b_3c_1 & a_5b_4c_1 & a_5b_1c_2 & a_5b_2c_2 & a_5b_3c_2 & a_5b_4c_2 & a_5b_1c_3 & a_5b_2c_3 & a_5b_3c_3 & a_5b_4c_3 \end{array}\right]$$

(A.6)

Definition A.26. (Multiple Product) The multiple product of N matrices, all having the same number of columns R, is the sum of their R N-ads, specifically,

$$\otimes\left(\mathbf{A}_{a\times R}, \mathbf{B}_{b\times R}, \mathbf{C}_{c\times R}, \cdots\right) \equiv \sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r \otimes \cdots$$

is the multiple product of $\mathbf{A}_{a\times R}, \mathbf{B}_{b\times R}, \mathbf{C}_{c\times R}, \cdots$.

Definition A.27. (*N*-adic Decomposition) For the *N*-array, **X**, the *N*-adic decomposition is the set of matrices {**A**,**B**,**C**,...}, each having *R* columns, whose multiple product is equal to **X**, i.e.,

$$\mathbf{X} = \otimes\left(\mathbf{A}_{a\times R},\quad \mathbf{B}_{b\times R},\quad \mathbf{C}_{c\times R},\cdots\right).$$

Definition A.28. (Array Rank) The rank of an *N*-array, **X**, is the minimum value of *R* (see Definition A.27) for which an *N*-adic decomposition of **X** exists, i.e., the minimum number of *N*-ads which is equal to **X**. The array rank of **X** is indicated by $R\ \ (\mathbf{X})$.

Definition A.28. Note 1. By definition, all *N*-ads are rank 1 arrays.

Definition A.29. (Kronecker Product) The Kronecker product (also called the *direct product*) of two matrices is formed by generating the product of each element of one matrix with each element of the other matrix in an ordered way. The Kronecker product of $\mathbf{A}_{n\times k}$ and $\mathbf{B}_{m\times p}$ is $\mathbf{C}_{nm\times kp}$, defined as:

$$\mathbf{C} = \mathbf{A}\otimes\mathbf{B} \equiv \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1k}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2k}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{nk}\mathbf{B} \end{bmatrix}$$

Definition A.30. (Positive Definite Matrix) A symmetric matrix is positive definite if and only if all of its eigenvalues are greater than zero.

Definition A.31. (Positive Semidefinite Matrix) A symmetric matrix is positive semidefinite if and only if all of its eigenvalues are greater than or equal to zero.

Definition A.31. Note 1. Positive semidefinite matrices are also called nonnegative definite matrices.

Definition A.32. (Gramian Matrix or Gramian Association Matrix) A positive semidefinite matrix.

Definition A.32. Note 1. There are four types of Gramian matrices. These are formed from the data matrix, **A**, by taking $\mathbf{A}'\mathbf{A}$ or $\mathbf{A}\mathbf{A}'$ after any preprocessing:

(1) Inner product matrix	(Data unstandardized)
(2) Cosine matrix	(Data scaled to unit variance)
(3) Covariance matrix	(Data centered about mean)
(4) Correlation matrix	(Data standardized)

See Chapter 3 for descriptions of centering, scaling, and standardizing.

APPENDIX B. THREE-MODE MODELS

Tucker Models, Three-Mode Factor Analysis, and Trilinear Models

Analysis of *N*-arrays was pioneered by Tucker[82] who developed the unfolding methodology for *N*-arrays discussed in Appendix A. His notation and methodology has been the backbone of multimode analysis ever since. Therefore, a brief explanation of his methodology is warranted. The discussion will begin with the 2-array notation and then progress to the more complex 3-array notation.

B.1 Tucker Models

Tucker's notation involves using pre-subscripts and post-subscripts around the letter representing the matrix to designate the mode of the row and column information, respectively. Choosing the lower case italic letters *i* and *j* to represent the two modes, a two-mode matrix (2-array) oriented *i* by *j* would be written ${}_i\mathbf{X}_j$. The subscript has several related but distinct roles: (1) as a general identification of the mode, (2) as a subscript identifying the mode to which an element belongs, and (3) as a variable identification symbol for the elements in the mode. For illustrative purposes, let ${}_i\mathbf{X}_j$ be a matrix of fluorescence intensities for a mixture of fluorophores measured at different excitation and emission wavelengths, i.e., an EEM.

The first usage of this notation would be to designate mode *i* as fluorescence excitation wavelengths and mode *j* as fluorescence emission wavelengths. The second usage is illustrated by assigning identification symbols to the different excitation wavelengths, i.e., 1_i, 2_i, 3_i, $\cdots$, N_i. The numeral in the identification symbol is the index value of the element, and the subscript indicates the mode. So, if one measured fluorescence intensities as a function of excitation at 10 nm increments from 250 nm to 350 nm, 3_i would represent 270 nm and as an index for any intensity at 270 nm. The number of elements in mode *i* is given by N_i and in mode *j* by N_j. Thus, the dimension of ${}_i\mathbf{X}_j$ is N_i by N_j. The third use of the mode symbol is as a variable identifier of a scalar element of the 2-array, e.g., x_{ij}.

Using this notation, the transpose of the 2-array , ${}_i\mathbf{X}_j$, is simply written as ${}_j\mathbf{X}_i$. Writing the product of two 2-arrays joins the post-subscript of the left 2-array with the pre-subscript of the right 2-array, since they must be identical, e.g.,

$${}_i\mathbf{A}_j\mathbf{B}_k \equiv \left({}_i\mathbf{A}_j\right)\left({}_j\mathbf{B}_k\right) \text{ or.}$$

$${}_i\mathbf{A}_j\mathbf{B}_k = {}_i\mathbf{C}_k$$

Now consider the factor analysis of the EEM, ${}_i\mathbf{X}_j$.

Factor analyzing the EEM, ${}_i\mathbf{X}_j$, into three matrices can be expressed as:

$${}_i\mathbf{X}_j = {}_i\mathbf{A}_n\mathbf{G}_n\mathbf{B}_j + \mathbf{N} \qquad \text{(B.1)}$$

where ${}_i\mathbf{A}_n$, is an N_i by N_n orthonormal factor matrix whose columns are the eigenvectors of ${}_i\mathbf{X}_j\mathbf{X}_i$, ${}_j\mathrm{B}_n$ is an N_j by N_n orthonormal factor matrix whose columns are the eigenvectors of ${}_j\mathbf{X}_i\mathbf{X}_j$, ${}_n\mathbf{G}_n$ is a diagonal matrix of square roots of eigenvalues associated with the Gramian inner product matrices ${}_i\mathbf{X}_j\mathbf{X}_i$ and ${}_j\mathbf{X}_i\mathbf{X}_j$, also called the "*core*" matrix, and **N** is a measurement error matrix of dimension N_i by N_j. This is equivalent to the singular value decomposition of ${}_i\mathbf{X}_j$, i.e.,

$$\mathbf{X}=\mathbf{USV}'. \tag{B.2}$$

The number of singular vectors, N_n, is chosen to equal the rank of ${}_i\mathbf{X}_j$. Therefore, after selecting only the first N_n singular vectors and values,

$${}_i\mathbf{A}_n\mathbf{G}_n\mathbf{B}_j={}_i\mathbf{U}_n\mathbf{S}_n\mathbf{V}_j. \tag{B.3}$$

Physically meaningful factors are obtained by coordinate rotation of the eigenvectors by a square transformation (or rotation) matrix $\mathbf{T}$. Letting ${}_i\mathbf{F}_n={}_i\mathbf{A}_n\mathbf{G}_n$, then,

$${}_i\mathbf{F}_{n*}={}_i\mathbf{F}_n\mathbf{T}_{n*} \tag{B.4}$$

$${}_j\mathbf{B}_{n*}={}_j\mathbf{B}_n\mathbf{T}_{n*}^{-1} \tag{B.5}$$

where $n*$ is the transformed derivational mode and ${}_n\mathbf{T}_{n*}\mathbf{T}_n^{-1}=\mathbf{I}$.

The extension into the three-mode case is straightforward; however, it involves an adjustment to the notation to accommodate higher order arrays. Consider a 3-array with modes *i*, *j*, and *k*, **X**. The elements in the array, x_{ijk}, form a box, which has the dimensions N_i, N_j, by N_k. Now, let the array be unfolded so that it now is a matrix with elements ordered *i* by (*jk*), where (*jk*) is read as *j*-outer loop, *k*-inner loop. Such an arrangement is written ${}_i\mathbf{X}_{(jk)}$. For example, let ${}_i\mathbf{X}_{(jk)}$ be a $N_i=5$ by $N_j=4$ by $N_k=3$, 3-array, such that

$${}_i\mathbf{X}_{(jk)}=\left[\begin{array}{ccc|ccc|ccc|ccc}
x_{111} & x_{112} & x_{113} & x_{121} & x_{122} & x_{123} & x_{131} & x_{132} & x_{133} & x_{141} & x_{142} & x_{143} \\
x_{211} & x_{212} & x_{213} & x_{221} & x_{222} & x_{223} & x_{231} & x_{232} & x_{233} & x_{241} & x_{242} & x_{243} \\
x_{311} & x_{312} & x_{313} & x_{321} & x_{322} & x_{223} & x_{331} & x_{332} & x_{333} & x_{341} & x_{342} & x_{343} \\
x_{411} & x_{412} & x_{413} & x_{421} & x_{422} & x_{223} & x_{431} & x_{432} & x_{433} & x_{441} & x_{442} & x_{443} \\
x_{511} & x_{512} & x_{513} & x_{521} & x_{522} & x_{223} & x_{531} & x_{532} & x_{533} & x_{541} & x_{542} & x_{543}
\end{array}\right]. \tag{B.6}$$

The notion of *j*-outer loop, *k*-inner loop is easily visualized by the following diagram:

$$
{}_i\mathbf{X}_{jk} = \begin{array}{cccccccccccccc}
 & \multicolumn{3}{c}{1_j} & \multicolumn{3}{c}{2_j} & \multicolumn{3}{c}{3_j} & \multicolumn{3}{c}{4_j} \\
 & 1_k & 2_k & 3_k & 1_k & 2_k & 3_k & 1_k & 2_k & 3_k & 1_k & 2_k & 3_k \\
\hline
1_i & & & & & & & & & & & & \\
2_i & & & & & & & & & & & & \\
3_i & & & & & & & x_{ijk} & & & & & \\
4_i & & & & & & & & & & & & \\
5_i & & & & & & & & & & & &
\end{array}
$$

Based on the discussion thus far, note that

1) ${}_i\mathbf{X}_{(jk)} \neq {}_i\mathbf{X}_{(kj)}$

2) The unfolding method illustrated in Appendix A represents an *i* by (*kj*) arrangement (considering *i*, *k*, and *j* to represent the **a**, **b**, and **c** modes, respectively). Two other important array orientations are presented below. These will be used in the factor analysis in this thesis:

$$
{}_j\mathbf{X}_{(ik)} = \left[\begin{array}{ccc|ccc|ccc|ccc|ccc}
x_{111} & x_{112} & x_{113} & x_{211} & x_{212} & x_{213} & x_{311} & x_{312} & x_{313} & x_{411} & x_{412} & x_{413} & x_{511} & x_{512} & x_{513} \\
x_{121} & x_{122} & x_{123} & x_{221} & x_{222} & x_{223} & x_{321} & x_{322} & x_{323} & x_{421} & x_{422} & x_{423} & x_{521} & x_{522} & x_{523} \\
x_{131} & x_{132} & x_{133} & x_{231} & x_{232} & x_{233} & x_{331} & x_{332} & x_{333} & x_{431} & x_{432} & x_{433} & x_{531} & x_{532} & x_{533} \\
x_{141} & x_{142} & x_{143} & x_{241} & x_{242} & x_{243} & x_{341} & x_{342} & x_{343} & x_{441} & x_{442} & x_{443} & x_{541} & x_{542} & x_{543}
\end{array}\right] \qquad (B.7)
$$

$$
{}_k\mathbf{X}_{(ij)} =
$$

$$
\left[\begin{array}{cccc|cccc|cccc|cccc|cccc}
x_{111} & x_{121} & x_{131} & x_{141} & x_{211} & x_{221} & x_{231} & x_{241} & x_{311} & x_{321} & x_{331} & x_{341} & x_{411} & x_{421} & x_{431} & x_{441} & x_{511} & x_{521} & x_{531} & x_{541} \\
x_{112} & x_{122} & x_{132} & x_{142} & x_{212} & x_{222} & x_{232} & x_{242} & x_{312} & x_{322} & x_{332} & x_{342} & x_{412} & x_{422} & x_{432} & x_{442} & x_{512} & x_{522} & x_{532} & x_{542} \\
x_{113} & x_{123} & x_{133} & x_{143} & x_{213} & x_{223} & x_{233} & x_{243} & x_{313} & x_{323} & x_{333} & x_{343} & x_{413} & x_{423} & x_{433} & x_{443} & x_{513} & x_{523} & x_{533} & x_{543}
\end{array}\right]
$$

(B.8)

B.2 Three-Mode Factor Analysis

Three-mode factor analysis involves decomposing the 3-array into orthonormal factor matrices and a core array. In summational form, the array is written:

$$
x_{ijk} = \sum_m \sum_p \sum_q a_{im} \; b_{jp} \; c_{kq} \; g_{mpq} + n_{ijk} \qquad (B.9)
$$

where *a*, *b*, and *c* are the elements of the three factor matrices and *g* is an element of the three-mode core matrix. In matrix notation,

$${}_i\mathbf{X}_{(jk)}={}_i\mathbf{A}_m\mathbf{G}_{(pq)}({}_p\mathbf{B}_j\otimes{}_q\mathbf{C}_k)+\mathbf{N} \qquad \text{(B.10)}$$

where the notation is the same as the 2-array case except for ${}_m\mathbf{G}_{(pq)}$, which is an unfolded core 3-array. The matrices ${}_i\mathbf{A}_m$, ${}_j\mathbf{B}_p$, and ${}_k\mathbf{C}_q$ are extracted from the Gramian inner product matrices,

$${}_i\mathbf{M}_i={}_i\mathbf{X}_{(jk)}\mathbf{X}_i \qquad \text{(B.11)}$$

$${}_j\mathbf{P}_j={}_j\mathbf{X}_{(ik)}\mathbf{X}_j \qquad \text{(B.12)}$$

$${}_k\mathbf{Q}_k={}_k\mathbf{X}_{(ij)}\mathbf{X}_k \qquad \text{(B.13)}$$

respectively, by solving,

$${}_i\mathbf{M}_i\mathbf{A}_m={}_i\mathbf{A}_m\ \ \lambda_m \qquad \text{(B.14)}$$

$${}_j\mathbf{P}_j\mathbf{B}_p={}_j\mathbf{B}_p\ \ \mu_p \qquad \text{(B.15)}$$

$${}_k\mathbf{Q}_k\mathbf{C}_q={}_k\mathbf{C}_q\ \ \nu_q \qquad \text{(B.16)}$$

where λ_m, μ_m, and ν_m are the eigenvalue matrices associated with ${}_i\mathbf{A}_m$, ${}_j\mathbf{B}_p$, and ${}_k\mathbf{C}_q$, respectively. N_m significant eigenvectors of ${}_i\mathbf{M}_i$ are used to form ${}_i\mathbf{A}_m$, N_p significant eigenvectors of ${}_i\mathbf{P}_i$ are used to form ${}_i\mathbf{B}_m$, and N_q significant eigenvectors of ${}_i\mathbf{Q}_i$ are used to form ${}_i\mathbf{C}_m$. The core array is then given by

$${}_m\mathbf{G}_{(pq)}={}_m\mathbf{A}_i^+\mathbf{X}_{(jk)}({}_j\mathbf{B}_p^+\otimes{}_k\mathbf{C}_q^+) \qquad \text{(B.17)}$$

where, ${}_m\mathbf{A}_i^+$ =$({}_i\mathbf{A}_m)^+$, etc. Since ${}_i\mathbf{A}_m$, ${}_j\mathbf{B}_p$, and ${}_q\mathbf{C}_k$ are all column-wise sections of column-wise orthonormal matrices, equation B.17 becomes simply

$$ {}_m\mathbf{G}_{(pq)} = {}_m\mathbf{A}_i\mathbf{X}_{(jk)}({}_j\mathbf{B}_p \otimes {}_k\mathbf{C}_q). \qquad \text{(B.18)} $$

Tucker's approach is only exact if all of the eigenvectors of the Gramian matrices are used, i.e., if $i=m$, $j=p$, and $k=q$. Tucker noted that "these procedures do not produce a least-squares approximation to the data." Kroonenberg and De Leeuw[83] corrected this deficiency by developing an alternating least squares (ALS) algorithm which they called TUCKALS3.

The TUCKALS3 solves iteratively for values of ${}_i\mathbf{A}_m$, ${}_j\mathbf{B}_p$, ${}_k\mathbf{C}_q$ and ${}_m\mathbf{G}_{(pq)}$ in turn, using the new estimates for these matrices in subsequent steps.

A Substep:

$$ {}_i\mathbf{M}_i = {}_i\mathbf{X}_{(jk)}({}_j\mathbf{B}_p\mathbf{B}_j \otimes {}_k\mathbf{C}_q\mathbf{C}_k)_{(jk)}\mathbf{X}_i \qquad \text{(B.19)} $$

$$ {}_i\hat{\mathbf{A}}_m = {}_i\mathbf{M}_i\mathbf{A}_m({}_m\mathbf{A}_i\mathbf{M}_i\mathbf{M}_i\mathbf{A}_m)^{-1/2} \qquad \text{(B.20)} $$

B Substep:

$$ {}_j\mathbf{P}_j = {}_j\mathbf{X}_{(ik)}({}_i\hat{\mathbf{A}}_m\hat{\mathbf{A}}_i \otimes {}_k\mathbf{C}_q\mathbf{C}_k)_{(ik)}\mathbf{X}_j \qquad \text{(B.21)} $$

$$ {}_j\hat{\mathbf{B}}_p = {}_j\mathbf{P}_j\mathbf{B}_p({}_p\mathbf{B}_j\mathbf{P}_j\mathbf{P}_j\mathbf{B}_p)^{-1/2} \qquad \text{(B.22)} $$

C Substep:

$$ {}_k\mathbf{Q}_k = {}_k\mathbf{X}_{(ij)}({}_i\hat{\mathbf{A}}_m\hat{\mathbf{A}}_i \otimes {}_j\hat{\mathbf{B}}_p\hat{\mathbf{B}}_j)_{(ij)}\mathbf{X}_k \qquad \text{(B.23)} $$

$$ {}_k\hat{\mathbf{C}}_q = {}_k\mathbf{Q}_k\mathbf{C}_q({}_q\mathbf{C}_k\mathbf{Q}_k\mathbf{Q}_k\mathbf{C}_q)^{-1/2} \qquad \text{(B.24)} $$

Starting values of ${}_i\mathbf{A}_m$, ${}_j\mathbf{B}_p$, and ${}_k\mathbf{C}_q$ are obtained from Tucker's method in Equations B.12-B.14. After evaluating the substeps, the core array, ${}_m\hat{\mathbf{G}}_{(pq)}$, is estimated using Equation B.16, completing an iteration. After each iteration, the fit is evaluated by examining the residuals of the data and the fit. Termination is based on reaching a minimum of the square of the norm of the residual array, *f*, i.e.,

$$f = \left\| {}_i\mathbf{X}_{(jk)} - {}_i\hat{\mathbf{A}}_m \hat{\mathbf{G}}_{(pq)} ({}_p\hat{\mathbf{B}}_j \otimes {}_q\hat{\mathbf{C}}_k) \right\|^2 = \left\| {}_i\mathbf{X}_{(jk)} \right\|^2 - \left\| {}_i\hat{\mathbf{A}}_m \hat{\mathbf{G}}_{(pq)} ({}_p\hat{\mathbf{B}}_j \otimes {}_q\hat{\mathbf{C}}_k) \right\|^2 .$$

(B.25)

This may also be written,

$$f = \left\| {}_i\mathbf{X}_{(jk)} \right\|^2 - \left\| {}_m\hat{\mathbf{G}}_{(pq)} \right\|^2 . \qquad \text{(B.26)}$$

$\left\| {}_m\hat{\mathbf{G}}_{(pq)} \right\|^2$ is a measure of the amount of the system variance explained by the model. The squares of the elements of the core array, g_{mpq}^2, indicate the amount of variance explained by the corresponding combination of factors (*m*, *p*, and *q*) from each mode.

Transformation of the three-mode principal factors is performed using three transformation matrices, ${}_m\mathbf{T}_{m*}$, ${}_p\mathbf{T}_{p*}$, and ${}_q\mathbf{T}_{q*}$.

$${}_i\mathbf{A}_{m*} = {}_i\mathbf{A}_m\mathbf{T}_{m*} \qquad \text{(B.27)}$$

$${}_j\mathbf{B}_{p*} = {}_j\mathbf{B}_p\mathbf{T}_{p*} \qquad \text{(B.28)}$$

$${}_j\mathbf{C}_{p*} = {}_j\mathbf{C}_q\mathbf{T}_{q*} \qquad \text{(B.29)}$$

The transformed core matrix is then

$${}_{m*}\mathbf{G}_{(p*q*)} = {}_{m*}\mathbf{T}_m^{-1}\mathbf{G}_{(pq)}({}_p\mathbf{T}_{p*}^{-1} \otimes {}_q\mathbf{T}_{q*}^{-1}) . \qquad \text{(B.30)}$$

It is possible to manipulate the core matrix in three-mode analysis to create a more parsimonious model. This may be done by setting small values of g_{mpq} to zero, during or after three-mode analysis, or by using chemical knowledge of the system at hand to select elements of the core matrix which are nonzero and setting the rest to zero. A special case of the restricted Tucker model is the trilinear model.

B.2 Trilinear Models

The trilinear model has core modes (*m*, *p*, and *q*) of equal dimension, so $N_m = N_p = N_q$, and the core array is *superdiagonal*, i.e., the *i*th diagonal element of the *i*th slab of the core array, is nonzero; and all other elements are zero. For example, the unfolded core array,

$${}_m\mathbf{G}_{(pq)} = \left[\begin{array}{ccc|ccc|ccc} a & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c \end{array}\right] \qquad \text{(B.31)}$$

with $a \neq 0$ and/or $b \neq 0$ and/or $c \neq 0$ is superdiagonal. In trilinear analysis, one seeks to find the minimal decomposition of a 3-array, i.e., the smallest value of N_m needed to fit the data or to find $N_m = R\left({}_i\mathbf{X}_{jk}\right)$. The core array is usually not utilized explicitly when performing trilinear analysis, rather the model is represented as an *N*-adic decomposition, thus,

$$\mathbf{X} = \otimes(\mathbf{A},\mathbf{B},\mathbf{C}) + \mathbf{N}. \qquad \text{(B.32)}$$

Solutions for **A**, **B**, and **C** are obtained by an ALS scheme, specifically, three-mode alternating least squares (3M-ALS), also known as PARAFAC. Setting the superdiagonal elements of the superdiagonal core array equal to one reduces equation B.10 to equation B.32. The algorithm proceeds as follows:

A Substep:

$${}_m\mathbf{M}_{(jk)} = {}_m\mathbf{G}_{(pq)}\left({}_p\mathbf{B}_j \otimes_q \mathbf{C}_k\right) \qquad \text{(B.33)}$$

$${}_m\hat{\mathbf{A}}_i = {}_m\mathbf{M}^{+}_{(jk)}\mathbf{X}_i \qquad \text{(B.34)}$$

B Substep:

$${}_p\mathbf{P}_{(ik)} = {}_p\mathbf{G}_{(mq)}({}_m\hat{\mathbf{A}}_i \otimes_q \mathbf{C}_k) \tag{B.35}$$

$${}_p\hat{\mathbf{B}}_j = {}_p\mathbf{P}^{+}_{(ik)}\mathbf{X}_j \tag{B.36}$$

C Substep:

$${}_q\mathbf{Q}_{(ij)} = {}_q\mathbf{G}_{(mp)}({}_m\hat{\mathbf{A}}_i \otimes_p \hat{\mathbf{B}}_j) \tag{B.37}$$

$${}_q\hat{\mathbf{C}}_k = {}_q\mathbf{Q}^{+}_{(ij)}\mathbf{X}_k \tag{B.38}$$

Iteration is complete when a minimum in f,

$$f = \left\|\mathbf{X} - \left(\otimes(\mathbf{A},\mathbf{B},\mathbf{C})\right)\right\|^2 \tag{B.39}$$

is attained. The trilinear model has the very desirable property that its decompositions are *rotationally unique*, under certain assumptions, vide infra. This implies that any minimal decomposition of the 3-array **X** will generate equivalent decompositions, i.e., decompositions which are identical except for scaling and permutation ambiguities.

To explain the conditions under which rotational uniqueness arises, some definitions are necessary.

Definition B.1. (Universal *k*-column independence) If a matrix **A** has j linearly independent columns, then **A** is said to have column rank j. This does not imply that all possible sets of j columns of **A** are linearly independent. However, if every set of k columns of **A** is linearly independent, then **A** is said to have universal k-column independence.

Definition B.1. Note 1. All matrices have universal 0-column independence. If the matrix has no 0 columns, then it also has universal 1-column independence. In addition, if the matrix has no 0 columns and has no identical or proportional columns, then it has universal 2-column independence.

Definition B.2. (*k*-rank) The maximum value of *k* for which a matrix has universal *k*-column independence is called the *k*-rank.

A corollary to a theorem of Kruskal,[12] which follows, states the condition required for rotational uniqueness. Let J_A, J_B, and J_C be the *k*-ranks of the three *R*-columns matrices **A**, **B**, and **C**, respectively. If

$$J_A + J_B + J_C \geq 2R + 2 \tag{B.40}$$

then $\otimes(\mathbf{A},\mathbf{B},\mathbf{C})$ is rotationally unique.

APPENDIX C. EIGENANALYSIS-BASED PROCEDURES

Direct Trilinear Decomposition

This is a capsule review of the eigenanalysis-based procedures (EBP) for direct trilinear decomposition obtained from Reference 10. References 36 and 62 contain more detailed derivations of the method.

Consider the simplest trilinear composed of two bilinear matrices, say EEMs of mixtures, each containing differing amounts of the same fluorophores. The trilinear model for this system is

$$\mathbf{X} = \otimes(\mathbf{A}, \mathbf{B}, \mathbf{C}) \tag{C.1}$$

where **X** is an *i* by *j* by *k* 3-array with $k = 2$ and **A**, **B**, and **C** are *i*, *j*, and *k* by *r*, respectively. Also assume that all elements in **C** are nonzero.

The two *i* by *j* slices of **X** may be written as:

$$\mathbf{X}_1 = \mathbf{A}\mathbf{D}_1\mathbf{B}' \tag{C.2}$$

$$\mathbf{X}_2 = \mathbf{A}\mathbf{D}_2\mathbf{B}' \tag{C.3}$$

where $\mathbf{D}_1$ is a diagonal matrix whose diagonal elements are the first column of **C**, and $\mathbf{D}_2$ is a diagonal matrix whose diagonal elements are the second column of **C**. Both $\mathbf{X}_1$ and $\mathbf{X}_2$ share common column and row spaces whose eigenvectors are **U** and **V**, respectively. These can, in turn, be related to **A** and **B** by transformation matrices **P** and **Q**, respectively, i.e.,

$$\mathbf{A} = \mathbf{U}\mathbf{P} \tag{C.4}$$

$$\mathbf{B} = \mathbf{V}\mathbf{Q} \tag{C.5}$$

Since it is a simple matter to compute the eigenvector matrices **U** and **V,** the task at hand is to compute the correct transformation matrices **P** and **Q**.

To solve for **P** and **Q**, let

$$\mathbf{L}_1 = \mathbf{U}'\mathbf{X}_1\mathbf{V} \tag{C.6}$$

and

$$\mathbf{L}_2 = \mathbf{U}'\mathbf{X}_2\mathbf{V} \tag{C.7}$$

Combining equations C.2, C.4, and C.6,

$$\mathbf{L}_1 = \mathbf{U}'\mathbf{X}_1\mathbf{V} = \mathbf{U}'\mathbf{A}\mathbf{D}_1\mathbf{B}'\mathbf{V} = \mathbf{U}'\mathbf{U}\mathbf{P}\mathbf{D}_1\mathbf{Q}'\mathbf{V}'\mathbf{V} = \mathbf{P}\mathbf{D}_1\mathbf{Q}' \tag{C.8}$$

and equations C.3, C.5, and C.7,

$$\mathbf{L}_2 = \mathbf{U}'\mathbf{X}_2\mathbf{V} = \mathbf{U}'\mathbf{A}\mathbf{D}_2\mathbf{B}'\mathbf{V} = \mathbf{U}'\mathbf{U}\mathbf{P}\mathbf{D}_2\mathbf{Q}'\mathbf{V}'\mathbf{V} = \mathbf{P}\mathbf{D}_2\mathbf{Q}' \tag{C.9}$$

Taking the product of $\mathbf{L}_1\mathbf{L}_2^{-1}$ yields

$$\mathbf{L}_1\mathbf{L}_2^{-1} = \mathbf{P}\mathbf{D}_1\mathbf{Q}'\mathbf{Q}'^{-1}\mathbf{D}_2^{-1}\mathbf{P}^{-1} = \mathbf{P}\mathbf{D}_1\mathbf{D}_2^{-1}\mathbf{P}^{-1} \tag{C.10}$$

and the product of $\mathbf{L}_1'\mathbf{L}_2'^{-1}$, gives

$$\mathbf{L}_1'\mathbf{L}_2'^{-1} = \mathbf{Q}\mathbf{D}_1'\mathbf{P}'\mathbf{P}'^{-1}\mathbf{D}_2'^{-1}\mathbf{Q}^{-1} = \mathbf{Q}\mathbf{D}_1\mathbf{D}_2^{-1}\mathbf{Q}^{-1} \tag{C.11}$$

Now,

$$\mathbf{L}_1\mathbf{L}_2^{-1}\mathbf{P} = \mathbf{P}\mathbf{D}_1\mathbf{D}_2^{-1} \tag{C.12}$$

and

$$\mathbf{L}_1'\mathbf{L}_2'^{-1}\mathbf{Q} = \mathbf{Q}\mathbf{D}_1\mathbf{D}_2^{-1} \tag{C.13}$$

which are both in the form of eigenvalue equations and can be solved directly for **P** and **Q**.

The extension to multiple slices in the third mode is accomplished by creating two matrices usually called G matrices, i.e., $\mathbf{G}_1$ and $\mathbf{G}_2$. The G matrices, for generalized matrices, are linear combinations of all of the slices in the 3-array such that all of the possible dyads are included for the analysis. The G matrices can be computed in a manner similar to $\mathbf{L}_1$ and $\mathbf{L}_2$ and are used in their place in the calculations.

After the estimates for **A** and **B** are determined, **C** is determined via least squares estimation as in 3M-ALS.